Learning new words: Effects of meaning, memory consolidation, and sleep

Jakke Johannes Tamminen

Submitted for the degree of Doctor of Philosophy

University of York

Department of Psychology

February 2010

Abstract

Although encountering novel words in one's own language in adulthood is not an uncommon event, the relevant cognitive processes have become the target of systematic investigation only in recent years. This thesis addressed three main questions regarding word learning. The first was concerned with the role of meaning: to what degree is meaning necessary in integrating new representations in the lexicon? Experiments 1-3 suggested that meaning is indeed important. In the absence of trained meaning novel words may "inherit" the meaning of neighbouring familiar words, possibly explaining some seemingly incompatible reports in the literature (Experiment 1). Experiment 3 showed that such inherited meaning is sufficient to allow integration of novel words in the lexicon. Having established the importance of meaning in lexical integration, the thesis moved to the second question: does knowledge of novel word meanings benefit from offline memory consolidation? Experiments 4-7 suggested that this is the case. Experiment 4 showed that consolidated novel words elicited faster semantic decisions than words learned just before testing, while Experiment 5 showed that cued recall of word forms is also enhanced over time. Experiments 6-7 refined these conclusions by using semantic priming paradigms, showing that novel word primes facilitate processing of semantically associated familiar words after a period of offline consolidation has been allowed to operate over an extended period of time involving several days and/or nights. The third question focused on the role of sleep in the consolidation of novel words: which aspects of sleep architecture are associated with lexical integration? Experiment 8 looked at sleep during the night after word learning and sought to clarify the roles sleep spindles and different sleep stages play in word learning. Spindle activity was associated with the emergence of lexical competition effects, suggesting that sleep has an active role in word learning, and that spindles in particular are associated with lexical integration. These effects were interpreted in light of complementary learning systems theories.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor Gareth Gaskell for his support and encouragement over the years, and for finding the perfect balance between helping me stay on track while at the same time allowing me space to grow as an independent scholar. I am also grateful to the members of my research committee, Silvia Gennari and Piers Cornelissen, for regularly providing a fresh outlook to my research. A similar invaluable service was provided by the psycholinguistics research group and all past and present members of the Gaskell lab group, to whom I extend my thanks.

I would also like to thank Robert Stickgold, Jessica Payne, and all other members of the Center for Sleep and Cognition at Harvard Medical School for generously hosting me during my overseas university visit and teaching me about the exciting possibilities of sleep research.

Many friends supported and helped me along the way; of these wonderful people I am particularly grateful to Vicky, Michelle, Liz, Mark, and Natalie. Claire, Gitte, and Roberto continue making C224 the best office in the department.

Maarit however deserves the greatest thanks for enduring the life of a PhD student's wife for more than three years, with the irregular hours and the unforgiving thesis submission deadline. I cannot imagine having managed it without her. Finally, thank you Matilda for coming along just at the right time when I needed to be reminded about the things that truly matter in life.

Declaration

This thesis contains original work completed solely by the author under the supervision of Professor Gareth Gaskell.

The research was supported by a studentship from the Economic and Social Research Council.

Data from this thesis were presented at the following conferences:

Experiment 3:

Tamminen, J., & Gaskell, G. *Novel words entering the mental lexicon: is meaning necessary for integration?* Poster presented at the EPS April meeting, Cambridge, 2008.

Experiments 4-6:

Tamminen, J., & Gaskell, G. *Changing dynamics in the mental lexicon: New lexical representations strengthen over time.* Poster presented at the 49th Annual Meeting of the Psychonomic Society, Chicago, November 2008, and at the EPS January meeting, London, 2009.

Experiments 5-8:

Tamminen, J., & Gaskell, G. *Offline consolidation facilitates access to novel word forms and meanings.* Talk given at EPS/CSBBCS Joint Conference, York, July 2009.

Experiment 8:

Tamminen, J., & Gaskell, G., Payne, J., Wamsley, E., & Stickgold, R. *Sleep spindle activity correlates with integration of newly learned words in the mental lexicon.* Poster presented at the 4th Computational Cognitive Neuroscience Conference, Boston, November 2009.

Chapter 1: From nonwords to novel words

1.1 Introduction

Learning new words is a skill most often associated with children. This is not surprising, considering the remarkable rate at which children seem to acquire new words. During a period starting at about 18 months, often referred to as the vocabulary spurt, children add up to nine words a day to their spoken vocabulary (Nazzi & Bertoncini, 2003, see also McMurray, 2007, for a recent view into the issue). While there is a vast literature concerning the cognitive foundations on which child word learning is based, less is known about word learning in adults. Adult word learning research has very much focused on second language learning (L2 learning). The subjective experience of L2 learning is often claimed to be slow and labour-intensive. This may tempt one to conclude that adults are not good word learners, at least not as good as children. While this may be the case to some extent, new L2 research suggests that novel words in a second language are learned very quickly. McLaughlin, Osterhout, and Kim (2004) tracked the neural correlates of word learning in L2 learners using event-related potentials (ERPs). They showed that only after about 14 hours of instruction in a foreign language, the N400, an ERP component thought to index semantic analysis, discriminated between nonwords (fictional made up words) and real words in the L2. It appears then that adults show more efficient learning of L2 words than subjective experience might suggest, at least when probed with a non-behavioural measure.

Much less is known about native language word learning though, yet word learning is a common event in adulthood. New words enter the language at a regular pace (often in connection with new technology, e.g., *blog*), and rare words occasionally return to common use at least temporarily (e.g., *redact* in the summer of 2009). Similarly, new terminology needs to be learned as we acquire new knowledge (e.g., *hippocampus* is likely to be a new word for a psychology undergraduate). However, there is now an emerging literature on various aspects of adult word learning, and a theoretical framework is beginning to emerge. In the following sections I will review the literature on adult word learning and highlight the various factors influencing the process.

1.2 Learning the form and meaning of novel words

What does adding a new entry to one's lexicon entail? To get a better idea of the process, it is useful to think about it in relation to existing models of word recognition (the work reported in this thesis deals with recognition rather than word production). Most current models of spoken word recognition (similar models exist in the written domain), such as TRACE (McClelland & Elman, 1986) and Shortlist (Norris, 1994), assume a pre-lexical level, consisting of phonemes and/or lower level items such as phonetic features. This level mediates between the heard raw signal and the lexical level, a collection of representations corresponding to words. In addition to a pre-lexical and lexical level there is also a semantic level associated with the meanings of the words. When a known word is heard, the signal will map onto the representations at the pre-lexical level, and activation will spread to the lexical level, resulting in the activation of a lexical representation and its semantics, terminating with word recognition. When an unknown, novel word is heard, activation of the pre-lexical level will take place as usual, but now there is nothing at the lexical level that corresponds to the signal. Hence for word learning to take place, at a minimum a new lexical representation will have to be established. In the next section I will discuss studies that can be taken as evidence for the creation of a new lexical representation.

1.2.1 Emergence of a new lexical representation

One of the early theoretical accounts of this process was introduced by Salasoo, Shiffrin, and Feustel (1985). These authors discussed codification, defined as "the development of a memory trace that responds as a single unit to a set of features and serves to label, code, name, or identify those features" (Salasoo et al, 1985, p. 51). As such, codification is simply another term for the creation of a new lexical representation. Using two threshold identification tasks, Salasoo et al. tracked the codification of visually presented novel words as a function of presentation duration, number of presentations, and time. The stimuli consisted of real words and meaningless novel words presented for short durations on the screen preceded (and followed in one variant of the task) by a mask of variable duration. The task was

simply to report the target word. While identification accuracy increased with the number of repetitions and increasing presentation times, there was initially a strong advantage for known words. This word advantage is usually explained by there being a lexical representation for known words, while no representation exists for nonwords. Interestingly, identification success for novel words converged with that of real words after only five repetitions. This signifies that codification has taken place, that is, a lexical representation had been established for the novel words. The strength of the newly created representation was evaluated by testing the same participants one year later. Real words and novel words were still identified equally accurately, despite an overall drop in performance. An explicit recognition task was also included where participants were asked to judge whether an item was included in the set presented a year earlier or not. Both real words and novel words were recognised above chance, with significantly better recognition performance with novel words than real words. These early data suggest that newly learned words develop a code (or a lexical representation) very quickly, and that this code is as strong and durable as that of real words.

A similar conclusion was reached by Monsell (1985). In a visual repetition priming experiment participants carried out lexical decision to real words and nonwords which were repeated three times over the course of the experiment, with the repetitions occurring with variable lags (i.e., the number of trials between two occurrences of a stimulus). The intertrial interval (ITI) was also manipulated, measured as the time between a response and the onset of a new trial and could be 500 ms or 1000 ms long. Repetition priming was found for real words across all ITIs, but with nonwords the effect was found only with the long ITI. The discovery of a repetition priming effect with nonwords was significant in its own right, but the effect of ITI was interpreted by Monsell as a potentially important variable in word learning. He argued that the nonword priming effect could be explained with the emergence of a new, possibly fragile, lexical unit. Furthermore, a new unit could only be established if enough time was allowed for learning to take place after the presentation of a stimulus. His data suggested that 500 ms was not enough for learning to take place, but the still remarkably short 1000 ms apparently was.

Forster (1985) examined the emergence of new lexical representations using masked visual repetition priming. A trial consisted of three visual stimuli: a visual mask, a brief prime (60 ms), and a target (e.g., ### – lock – LOCK). The typical

finding in this paradigm is that when the prime and target are the same word, processing of the target is faster than when the two are different words. Due to the very short duration of the prime, it is argued that priming effects reflect the repeated access to the target word representation, rather than the facilitation afforded by an episodic memory trace for example. In fact, participants tend to have no conscious awareness of the identity of the prime. This lexical interpretation is supported by the lack of priming in nonwords, which do not have lexical representations. This paradigm was seen by Forster as an ideal way of evaluating the emergence of new lexical representations, and an improvement over the task used by Salasoo et al. (1985) where the advantage of real and novel words over nonwords could have been the result of permanent episodic representations, as opposed to true lexical representations.

Participants were taught unfamiliar rare words (e.g., *pinery*) and their meanings. While these words showed no repetition priming effects in a lexical decision task before training, they did result in priming after training. No effect was found for word-like nonwords (defined here as nonwords which suggest a meaning, such as *bellowbag*), and a small priming effect was found surprisingly for nonwords (conventional nonwords, such as *tovit*) also. In another experiment participants studied meaningless novel words. The task was a standard recognition task where stimuli were classified as old (seen in training) or new (not seen in training). A priming effect in this task was found for the trained items. Does this suggest that providing the novel words with meaning offers no advantage in word learning? Forster argued that this is not the case, as the recognition task does not require lexical access. Together these experiments do however again suggest rapid emergence of new lexical representations.

Rajaram and Neely (1992) extended Forster's results. A masked repetition priming paradigm was again used with both lexical decision and an explicit recognition tasks. When participants were instructed to try to learn the words in the study list, masked priming effects were found for studied nonwords but not for unstudied nonwords. In line with earlier researchers Rajaram and Neely concluded from this that a temporary lexical entry had been created for the studied nonwords. They further postulated that this may have been the case because participants were explicitly asked to learn the nonwords. In their second experiment participants were not told to learn the nonwords, instead they were asked to read the study list and

decide whether an item's pronunciation sounded pleasant or not. Masked repetition priming was again found for the nonwords that occurred in the study list, but not for nonwords encountered for the first time. The priming effect in this experiment was smaller though than in the first one (although not statistically significantly smaller). Rajaram and Neely proposed that the nonwords had created temporary lexical entries and that the strength of the entries depended on the instructions given to participants in the study phase, with intentional learning leading to stronger entries.

Johnston, McKague, and Pratt (2004) used the same priming paradigm to see whether a novel phonological lexical representation also gives rise to an orthographic representation. Unfamiliar rare words and their meanings were trained in the auditory modality only. This was followed by a repetition priming task in the visual modality. If new lexical representations emerged as a consequence of the phonological training, and if a parallel orthographic representation also emerged, repetition priming effects should be detectable. This was indeed the case although it appeared that the orthographic representations were underspecified to some degree. Prime novel words that differed from the target by two letters also showed priming. This happened only with novel words, not with real words. There was no evidence of improvement of orthographic specificity with further three exposures to the orthography.

Ellis, Ferreira, Cathles-Hagan, Holt, Jarvis, and Barca (2009) identified two further phenomena that may be used as markers of lexical representation. Firstly, longer nonwords are read slower than shorter nonwords. This length effect is reduced for familiar words, suggesting that accessing a lexical representation allows for parallel processing (as opposed to serial letter-by-letter reading) of written words. Secondly, the length effect in real words is particularly reduced when words are presented in the right visual field, presumably because anything perceived in the right visual field has direct access to the language-dominant left cerebral hemisphere due to the anatomy of the visual neural pathways. In an experiment involving orthographic and semantic word training tasks over two days, Ellis et al. (2009) showed that novel words which prior to training showed the typical length effect in both visual fields became more word-like after training in that the length effect became attenuated in the right visual field compared to the left, coupled with a reduction in overall reading times and error rates. The authors argued that this signalled a change from serial to parallel processing particularly in the left mid-

fusiform gyrus, an area thought to be responsible for visual word recognition, indicating the generation of a new lexical representation.

Fast and efficient learning was observed in another test of word form knowledge that was employed by Leach and Samuel (2007). These authors used a threshold discrimination task where novel spoken words were presented in varying levels of white noise, and the participant's task was to report the heard word. Novel words were trained in various tasks across five experiments (these will be discussed in more detail later), but performance in the noise task tended to be good in all experiments. Accuracy ranged from about 70% to 80% on the first day of training, and increased to about 90% correct by the fifth day of training. Also, the level of noise at which recognition could take place increased over the five days.

Unlike the priming studies discussed above, the noise threshold task is a measure of explicit (or declarative) knowledge of word forms. The distinction between declarative and nondeclarative forms of memory is often highlighted in studies with amnesic patients. These patients tend to be impaired on declarative memory (e.g., recall and recognition performance), while relatively unimpaired with nondeclarative tasks. Musen and Squire (1991) argued that nondeclarative memory can be used in amnesic patients to support the acquisition of novel lexical information. Amnesic patients and normal controls were asked to read a list of real words and nonwords with some of the words and nonwords occurring only once in the list, and some repeated several times. Both patients and controls showed faster reading times to repeated nonwords as the number of repetitions increased. The authors did not compare statistically the reading times to words and nonwords, but judging by their figures it appears that by the end of the list repeated nonwords were read as quickly as repeated real words. This strikingly good performance was contrasted in a simple recognition test, where the patients performed significantly worse than controls. It appears then that even in amnesic patients in a nondeclarative measure performance with words and nonwords converges as exposure to the nonwords increases, reminiscent of the Salasoo et al. (1985) data. It is possible to speculate that new lexical representations for the nonwords emerged in this study both for amnesics and controls, but that these new representations were detectable only through a nondeclarative task in the amnesic group.

Interestingly, Duff, Hengst, Tranel, and Cohen (2006) have shown that amnesic patients can under certain circumstances learn labels for novel objects even

when measured through explicit means. In this experiment patients derived self-generated labels for visual shapes in a collaborative task with a partner. Although these patients were highly impaired in learning pre-determined labels for similar shapes in a control condition, they learned the self-generated labels nearly equally well as normal control participants, and still recalled 80% of the labels six months later (compared to 83% in the control group). The authors argued that amnesic patients are poor at learning arbitrary relations between labels and objects, but when the labels are self-generated they have a meaningful relation to the shapes, with the latter perhaps being a case of hippocampus-independent learning.

The studies discussed so far have shown emergence of lexical representations under fairly simple circumstances, where the participant's task has been merely to learn individually presented words. An example of a more demanding paradigm, and perhaps a more realistic learning situation, comes from the work on word segmentation. Saffran, Newport, and Aslin (1996) exposed adult participants to an artificial language in the spoken modality, consisting of six meaningless words (e.g., *babupu*). Participants were exposed to the words in a continuous stream of syllables where the only cue to the beginnings and endings of the words was provided by the probability with which the syllables occurred together (any acoustic cues were absent since the stimuli were presented by a speech synthesizer). Syllables within words occurred together frequently, while the final syllable of one word and the first syllable of another word occurred together only rarely. If participants were able to segment the stream into word units using these transitional probabilities, they should be able to discriminate the novel words in the artificial language from nonwords randomly generated from the same syllables. This was indeed the case, participants performed above chance in a 2-alternative forced choice (2AFC) task. They performed above chance also in a more difficult variant of the task, where the nonwords differed from the novel words by only one syllable. Other experiments by the same research group have shown that children as young as 8 months are able to segment the stream and learn the novel words (Saffran, Aslin, & Newport, 1996).

Dahan and Brent (1999) introduced another mechanism through which novel words can be segmented from continuous speech. They discussed the INCDROP (incremental distributional regularity optimization) model, which suggests that novel words are segmented from the speech stream by recognising familiar units (i.e., known words), extracting these from the stream, and treating any stimuli left over as

novel words. In a number of experiments Dahan and Brent exposed participants to spoken meaningless novel words in isolation (e.g., *dobu*) and embedded in longer utterances (e.g., *dobuneripo*). If the novel word was learned, the participant should be able to segment that new word out from the utterance, and treat the remaining portion of the utterance as another novel word (e.g., *neripo* extracted from *dobuneripo*). Lexical decision and recognition tasks showed that participants were able to segment the utterances successfully. The authors stated that this behaviour does not necessarily mean that a new lexical representation has been created for any of the novel words segmented from the utterances. While it is possible that a form-based lexical representation could cause these effects, a familiarity effect associated with an episodic memory trace is equally likely in this study, as in the Saffran et al. studies. A related criticism concerns the explicit nature of the tasks used in many of the segmentation studies, such as lexical decision, which requires a metalinguistic judgement.

One way to avoid asking participants to make explicit judgements is to monitor brain activity as novel stimuli are listened to. Sanders, Newport, and Neville (2002) measured ERPs to continuous novel word streams both before training on the words and after training. They were interested in seeing if training induces the emergence of the N100 component in response to novel word onsets. The N100 is known to occur at word onsets with real words, and should thus be an indicator of the novel words having generated lexical representations. The N100 amplitude to novel word onsets did indeed increase after training, but only in a group of participants who also showed successful word learning in a behavioural recognition test. An increased N400 to the novel words was also observed post-training in all participants. The authors suggested that the N100 indexed fast, online segmentation, while the N400 indexed lexical search strategies. Since these data were obtained using a paradigm that did not require an explicit metalinguistic judgement, and the effects included ERP components known to be related to lexical processing, it seems that the involvement of new lexical representations was likely.

1.2.2 Linking a new lexical representation with meaning

Many of the studies I have described so far used paradigms where participants' knowledge of the novel word forms was tested rather than their

knowledge of the word meanings. It seems that under these conditions a novel lexical representation can be created remarkably quickly. However, many studies have paired novel words with a meaning, for example a visually presented object, thus requiring learning of the word and its referent and tested for recall of both word forms and the meanings. The advantage of these studies is that they are more realistic in emulating the word learning taking place naturally, and also give some idea of whether learning the meaning of a novel word is as quick and efficient as learning the form.

Gupta (2003) reported data from two word learning experiments where the main focus of interest was in uncovering a correlation between word learning, nonword repetition, and immediate serial recall. Adult participants were asked to learn the names of imaginary animals (Experiment 1), and cartoon aliens (Experiment 2). Learning was measured by asking participants to name the pictures. Overall, participants learned the names successfully, with 78% of naming trials correct in Experiment 1, and 46% correct in Experiment 2. Gupta did not comment on the drop in performance in Experiment 2, but it may be due to the smaller number of training trials or the nature of the pictures (e.g., the aliens may have been less distinctive from each other). Nonetheless, participants were able to reliably learn most of the novel words and their referents. In addition, a correlation was found between word learning and nonword repetition, as well as word learning and immediate serial recall (similar correlations have been reported with children too). Gupta argued that accurate nonword repetition leads to more accurate word learning, hence the correlation. The correlation between learning and serial recall was explained by a correlation between serial recall and nonword repetition. Serial recall plays an important role in nonword repetition because for accurate repetition to take place, the sublexical units of a nonword must be repeated in the correct serial order.

Another paradigm using pairings of novel words and pictures was introduced by Breitenstein and Knecht (2002). Participants were shown line drawings of common objects together with a spoken novel word. The task was to indicate as quickly as possible whether the pair was “correct” or “incorrect”. The “correctness” of the pairings was determined by their statistical co-occurrence. Each novel word occurred with a specific drawing with high frequency, and with other drawings with low frequency. The challenge was to see if participants picked up on these statistical properties of the stimuli, and learned to associate the novel words with the drawings

they were most often associated with. Training on the stimuli took place during five sessions across five days. Performance accuracy increased from initial chance level to about 60% correct by the end of the first session, and further increased up to about 90% by the end of the fifth session. Two further sessions were added, one took place a week later, and another one month later. Performance remained at about 90% in these sessions, showing long-term learning of the meanings.

Nelson, Balass, and Perfetti (2005) used a word learning paradigm where rare real words, such as *clowder*, are used instead of artificially created novel words. One advantage of this method is that the novel words are more likely to conform to the phonological properties of the language than if they are artificially created. In a later section I will describe studies showing why this is an important consideration. Participants were trained on 35 novel words by presenting them together with a definition. Training continued until all of the word meanings had been learned (or until 2.5 hours of training had been completed). In the testing phase the trained words were presented with similar foils and participants' task was to decide whether the word was presented in the training phase (old word) or not (new word). The main question Nelson et al. were interested in was whether training modality affected learning, hence some of the novel words were presented visually in the training phase and others auditorily. Modality was also manipulated in the recognition task, to see if there is a recognition advantage for items which are presented in the same modality in both training and testing. The authors did not report whether any participants failed to reach the criterion in the training, so it seems reasonable to assume that all participants learned the 35 novel word meanings within the allowed time. Visually presented novel words required fewer training trials than auditorily presented words, and participants were more accurate in recognising the trained items when they were presented in the same modality in test as in training.

The three studies reviewed above suggest that adult learners are quickly and reliably able to pair a new word with its meaning. However, this merely tells us that people are able to memorise the pairings. It would be more interesting to see whether newly learned meaningful words give rise to implicit effects outside of the participant's conscious control, in the same way as familiar words do. For these kinds of effects we can look to neuroimaging studies. Perfetti, Wlotko, and Hart (2005) looked at word learning and ERPs. Participants were trained on rare words and their definitions, and asked to learn as many words as they could in a set time.

Learning was measured in a semantic decision task where the novel words (and fillers) were presented paired with a familiar real word that could be semantically related or unrelated to the novel one. The task was to decide whether the pair was related or not. High success rate in this task indicated good learning of the novel word meanings. A mean success rate of 82% was achieved with the novel words, which was comparable to the rate achieved with familiar words of medium frequency (87%). ERPs were recorded during the task, and showed a higher N400 in the unrelated condition, for both real words and novel words, but not for untrained novel words. The behavioural data together with the N400 finding suggest that people learned the words to a degree where their meanings were processed neurally in a similar way to familiar words.

Another ERP study trained participants on novel words from the artificial "Keki" language (McCandliss, Posner, & Givon, 1997). The training took place over 50 hours, where the Keki words and their meaning were taught using interactive computer tutorials. Testing, while ERPs were recorded, took the form of passive viewing, semantic judgement, and feature search tasks. ERP analysis was carried out on an early N100 window and a later P200 window. The N100 window was found to be sensitive to orthographic effects: consonant strings elicited a higher negativity than familiar English words. Trained and untrained Keki words elicited an intermediate negativity. This was unsurprising, as the Keki language orthography is similar but not identical to English orthography. In the P200 window results varied as a function of task and training. In the semantic task, while the mean amplitude for the trained Keki words reduced over the training sessions, both consonant strings and untrained Keki words remained fairly static, with significantly lower amplitudes to trained than untrained Keki words after 20 hours of training. There was no such difference in the passive viewing task and in the feature search task. These data again indicate that participants learned the meaning of the novel words.

1.3 Novel words' impact in the mental lexicon

The studies discussed in the previous section converge on a conclusion: people are good at learning the form and meaning of newly presented words, both in tasks measuring explicit and implicit knowledge. Salasoo et al. (1985) showed that novel words can acquire a lexical representation (a "code" in their terminology) even

under very impoverished training conditions that is robust over time, detectable even a year after its assumed creation. Perfetti et al. (2005) showed that people can learn the meanings of a fairly large number of novel words within a short time, and Breitenstein and Knecht (2002) showed that the link between a word form and its referent can be established by their statistical co-occurrence alone and is not weakened during several months of inactivity between training and re-test.

The evidence reviewed so far seems to suggest that a new lexical representation is set up quickly (perhaps even within 1000 ms as suggested by Monsell). Lexical representations however have many different properties and display many unique behaviours. Many of these phenomena have to do with the way lexical representations interact with each other and with other levels in the lexicon. Studies I have discussed so far do not tap into these more dynamic behaviours. The distinction between knowing the form or meaning of a word, and the word engaging with other words or linguistic representations in the lexicon is explicitly defined by Leach and Samuel (2007). They discuss "lexical configuration" and "lexical engagement". Lexical configuration refers to knowledge of the factual information about a word, such as its spelling or phonology, and the meaning of the word, and it is this type of knowledge that I have discussed so far. Lexical engagement on the other hand refers to the dynamic behaviour of the novel words with respect to other lexical or sublexical units. As pointed out by Leach and Samuel, semantic priming would be one example of lexical engagement, where exposure to one word (the prime, e.g., *doctor*) influences the processing of another word (the target, e.g., *nurse*), in this case by speeding up the processing of the target. In order to conclude that novel words have been entered in the mental lexicon, one would like to see data related not only to lexical configuration, but also evidence of lexical engagement, as this would show that a new lexical representation has been integrated into the lexicon, and has formed links with other lexical items and/or sublexical levels.

Another reason why finding evidence of lexical engagement is important has to do with the argument for episodic memory traces. It is difficult to reject this argument based on word form learning alone, as knowledge of novel word forms could conceivably be stored in episodic memory rather than in the lexicon[1]. Showing

[1] Although note that some theorists have proposed that the lexicon is episodic in nature (e.g., Goldinger, 1996).

that novel words engage in behaviours unique to lexical representations would be useful in evaluating the lexical status of novel words. Relevant evidence is now available in the domains of lexical competition, perceptual learning, and semantic priming. These will be discussed next.

1.3.1 Lexical competition

Spoken word recognition is a competitive process. As the spoken signal gradually unfolds, listeners entertain several hypotheses about the identity of the word (Marslen-Wilson, 1987). Hearing the first few sounds of a word will activate a cohort of word candidates matching the incoming signal. As the ensuing lexical competition continues and as more of the signal is heard, the number of matching candidates is reduced until only one remains. The point at which only one word matches the signal is known as the uniqueness point (UP). Many short words do not become unique until the end of the word, but most long words have a UP before the offset.

The UP of a word will affect the word's recognition time. Words with early UP are recognised faster than words with a late UP. In late-UP words several candidates will compete for recognition for longer than in early-UP words where competition is resolved early due to the quick exclusion of non-matching candidates. The effect of UP on word recognition times is now well documented in a range of response time experiments (see McQueen, 2007, for a brief review) and experiments monitoring the evaluation of word candidates using eye-tracking (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). Lexical competition is a good example of lexical engagement, an instance of clear interaction between lexical representations, and hence a potentially reliable test of "lexical integration" of novel words. In this thesis I will use the term lexical integration to refer to the process of a novel word integrating in the mental lexicon and interacting with familiar words and other levels of the lexical system, such as phonemic or semantic levels.

1.3.1.1 Novel words engage in lexical competition with each other

One can ask three questions about lexical competition and novel words: do novel words compete with each other, do novel word compete with existing words, and do existing words compete with novel words? Magnuson, Tanenhaus, Aslin, and

Dahan (2003) addressed the first question. Participants were trained on bisyllabic novel words (e.g., *pibo*), manipulating both the frequency of the novel words (by varying the number of times they were presented in training) and the frequency of their onset competitors (e.g., *pibu*) and rhyme competitors (e.g., *dibo*). Each novel word was associated with a visual shape, and testing consisted of trials where the participant was asked to click on a specific shape (e.g., *"click on the pibo"*). Eye-tracking was used to evaluate the activation of the target and competitors as a function of time. The artificial vocabulary made up of the newly learned words showed many of the same behaviours as real vocabularies do. High-frequency targets were fixated more than low-frequency targets, and the same applied to high- and low-frequency competitors. The critical finding showing lexical competition effects was that targets presented with low-frequency competitors were fixated more than targets presented with high-frequency competitors. Both onset and rhyme competitors received more fixations than phonologically unrelated competitors, although with training onset competitors showed an advantage over rhyme competitors, a pattern observed also with real words (Allopenna et al., 1998).

Having shown lexical competition effects within the artificial vocabulary, Magnuson et al. (2003) wished to see if existing words affect the processing of the vocabulary. To do this, they designed novel words that would fall into dense or sparse real-word phonological neighbourhoods. Real-word neighbourhood effects would indicate that existing words affect the processing of the novel vocabulary, whereas the absence of such effects would suggest that the novel vocabulary is isolated from the existing vocabulary. No reliable evidence of real-word intrusions was found, although there was a trend of neighbourhood effects with low-frequency novel words. These data then indicate that novel words do show word-like properties in terms in lexical engagement, but only as far as they engage with each other. This is encouraging but does not provide evidence for the novel word representations having formed links with existing items in the lexicon. A stronger test of lexical integration would be to show that novel words compete with existing words. Evidence of this is what I will discuss next.

1.3.1.2 Novel words engage in lexical competition with existing words

Gaskell and Dumay (2003) exposed participants in a phoneme monitoring task to meaningless spoken novel words, and sought to examine whether these novel

words engage in lexical competition with existing words. To this end, the novel words were derived from existing base words with relatively early uniqueness points (e.g., *cathedruke* derived from *cathedral*). Lexical competition effects were tested by measuring lexical decision times to the base words. The authors found that being exposed to *cathedruke* slowed down lexical decision times to *cathedral*, not immediately after exposure, but after three days of training. To avoid an explanation based on a response bias to the base words, the same results were shown in a pause detection task, where a short silent pause is embedded in a word. The participant is asked to decide whether a pause was present or not, and the time to make this decision is taken to be an indication of the level of lexical activity at the time (Mattys & Clark, 2002). Explicit form knowledge was tested in a 2AFC task requiring discrimination between trained novel words and similar sounding untrained foils. Highly accurate performance was found immediately after training, with only slight improvement with time.

The lexical decision and pause detection data showed that newly acquired words engage in lexical competition with existing words, and the 2AFC data showed that the form of the novel words had been acquired to a high degree. Together these data suggest that the novel words have been integrated in the mental lexicon. Furthermore, the dissociation between immediate form learning and delayed lexical integration is striking.

The time course of lexical integration was further narrowed down by Dumay, Gaskell, and Feng (2004) who showed that lexical competition effects can be observed, with pause detection, as soon as 24 hours after exposure. Again, the 2AFC task was almost at ceiling immediately after exposure. They also failed to detect any improved learning in a semantic training task over a purely phonological training task, suggesting that meaning is not obligatory for lexical integration. Tamminen and Gaskell (2008) extended the re-test of lexical competition effects showing that the effect, measured with lexical decision, is observable even 8 months after initial exposure. This was interpreted as evidence against an argument that the lexical representations created in these experiments might be situational or episodic in nature.

While the delay in lexical integration is now fairly well established, it remains unclear why the delay is necessary. One possibility is that it takes time for the new phonological detail to refine to a degree that allows lexical competition to

take place (Gaskell & Dumay, 2003). Another potential reason is that the delay reflects the operation of different memory systems. In some computational complementary learning systems (CLS) accounts new information is initially stored separately from existing information (presumed to correspond at the neural level to hippocampal structures), and interleaved and integrated over a longer time course with existing memories in neocortical areas where information is represented in a distributed fashion (e.g., McClelland, McNaughton, & O'Reilly, 1995). According to some researchers this interleaving takes place during sleep (e.g., Wilson & McNaughton, 1994). Hence it may not be simply the passage of time between exposure and test that is critical in lexical integration, but sleep. Dumay and Gaskell (2007) tested this account by exposing participants to novel words in the morning or in the evening, and testing for lexical competition effects immediately, 12 hours, and 24 hours later. Thus the evening group had had a night's sleep between exposure and the second test, while the morning group had not. Lexical competition effects, measured with pause detection, were absent immediately after exposure in both groups, and found only in the evening training group after 12 hours. The third test, 24 hours after exposure, when both groups had slept, showed lexical competition effects in both groups.

Dumay and Gaskell (2007) discussed three potential explanations for these data. First, it may be that lexical integration is associated with a certain circadian state, in which case sleep is irrelevant. Second, the critical factor may not be sleep, but an absence of potentially interfering stimulation. Third, sleep may provide an optimal brain state for the new lexical knowledge to be consolidated in long-term memory, as mentioned above with respect to CLS models. I will return to this question in Chapter 6, where I will present data relevant to the latter explanation. For now it suffices to state that it is clear that lexical integration, when measured by lexical competition, does not take place immediately after training, but only after some form of offline consolidation has taken place. Note that the term offline is used here, as it is typically used in this literature, to highlight the fact that consolidation occurs in the absence of further exposure to the materials to be consolidated.

One possible problem with the Dumay and Gaskell (2007) sleep data is that the test of lexical competition was repeated at three points in time. This leaves open the possibility that what was consolidated was procedural knowledge of the testing task, rather than novel lexical representations. To overcome this limitation Davis, Di

Betta, Macdonald, and Gaskell (2009) modified the procedure so that only one testing session was required. Participants learned one set of novel words on day 1, and another set on the following day. Testing took place immediately after the second training session on day 2. Hence one set of novel words had been learned on the previous day, and had had a chance to be consolidated prior to testing. As expected, the consolidated set showed lexical competition effects (measured with lexical decision to base words), while the unconsolidated set did not. A 2AFC task was also included, where highly accurate performance was observed for both sets, although the consolidated set had higher accuracy scores. A speeded vocal repetition test (shadowing) showed faster repetition times to consolidated novel words compared to untrained novel words, but this was not the case for unconsolidated items. Finally, participants were asked to rate the novel words for meaningfulness on a 7-point scale. Consolidated words were rated higher than unconsolidated words, although since no meaning was provided in training, it is hard to say on what basis the ratings were provided.

Davis et al. (2009) used the same training regime in an fMRI experiment. Here the primary measure of lexical integration was the neural response to consolidated novel words, unconsolidated novel words, untrained novel words, and real words. An elevated response to unconsolidated and untrained novel words compared with real words was found in a bilateral region of the superior temporal gyrus, and also in bilateral motor cortex, supplementary motor area, and cerebellum. Interestingly, this contrast was not found with the consolidated novel words. Further analyses revealed that the elevated response to novel words was reduced for consolidated novel words in bilateral motor and somatosensory areas, left premotor cortex, supplementary motor area, and the right cerebellum. Hence there was an interesting convergence of the behavioural and neural data: words that had had a chance to consolidate resulted in reduced shadowing latencies and reduced neural activation in regions associated with phonological processing. This was not found with words learned a few hours before testing, again supporting a CLS account.

Lexical competition is a process not limited to spoken word recognition. Many models of written word recognition postulate a similar mechanism where orthographic neighbours, such as *banish* and *vanish*, inhibit each other's activation. Hence orthographic neighbourhood size and the frequency of the neighbours should affect a written word's recognition time. Bowers, Davis, and Hanley (2005) showed

this to be the case by teaching participants novel words which, if lexically integrated, should become competitors with existing "hermit" words, that is, words with no prior neighbours (e.g. *banana* and the novel word *banara*). Novel words showed weak inhibitory effects immediately after training and strong effects one day later. This finding extends the data from spoken word recognition, and shows a similar time course, with weak immediate effects and strong effects one day later, with no exposure in between.

Following up the Bowers et al. (2005) study, Qiao, Forster, and Witzel (2009) used masked form priming as an alternative way to assess a novel word's influence on a familiar word. This experiment relied on the finding that priming between two similar (not identical as in repetition priming discussed earlier) written word forms depends on the lexicality of the prime, with nonword primes resulting in facilitation (faster RTs to targets preceded by a nonword prime, e.g., contrapt – CONTRACT) and word primes resulting in no priming (e.g., contrast – CONTRACT). Thus, if a newly learned word has been integrated in the mental lexicon, it should not result in facilitation. Qiao et al. (2009) tested this hypothesis using the same materials and training regime as Bowers et al. (2005), and found facilitation on both testing days. This led the authors to argue that the newly learned words were stored in episodic memory and had not been integrated in the lexicon, thus being unable to compete with familiar words. An episodic memory trace might have been able to result in inhibition in the Bowers study if seeing the base word *banana* activated the episodic trace of *banara,* initiating a postlexical orthographic check of higher intensity than before training.

Finally, the contrast between participants' explicit knowledge of the newly learned word forms on the one hand and lexical integration on the other was examined in a recent study by Fernandes, Kolinsky, and Ventura (2009). This study elegantly bridged the gap between the word segmentation studies described earlier and lexical competition as a measure of lexical integration. Recall that several authors have now demonstrated that adults and children are able to segment a spoken string of syllables into words based on the statistical pattern of co-occurrence between the syllables. This suggests that novel words have been extracted from the signal, and that at least rudimentary lexical representations have been set up for these words. This alone however does not allow us to conclude that the new words have been integrated in the mental lexicon. Fernandes et al. (2009) showed that this can

happen at least when there is more than one segmentation cue present, and when the cues do not contradict each other. When participants were exposed to an artificial language where the segmentation cues included transitional probabilities and word-like phonological structure, they later correctly discriminated novel words from part-words formed by incorrectly combining syllables from the artificial language. In addition, there was evidence of the novel words engaging in lexical competition, as manifested by slow responses in a lexical decision task to familiar base words that overlapped with the novel words. In another experiment where the two segmentation cues contradicted each other, participants appeared to segment the language based on the phonological word-likeness, but no lexical competition effect was found either immediately or one week later. Apparently only phonological segments extracted based on more than one consistent type of cue are established strongly enough to allow lexical integration to take place. Furthermore, in contrast to the work of Gaskell and colleagues, in the cases where the lexical competition effect was found, it was seen immediately after training as well as one week later, although the effect was significantly stronger in the delayed test. The immediate emergence of the lexical competition effect may have been due to the training task, which here took the form of a speech segmentation task. The segmentation requirement may have highlighted the overlap between the novel words and their base words, hence enabling faster linking between the two. It is also plausible that the segmentation task allows more incremental learning than presenting all novel words in isolation in one task, provided that the novel words were segmented from the speech stream in a gradual manner. This may have resulted in a degree of distributed learning as opposed to purely massed learning. Distributed learning has been shown to result in more robust memory traces than massed learning (e.g., Litman & Davachi, 2008), and this advantage may have at least partially compensated for the lack of offline consolidation in the immediate test session.

There is now a healthy number of studies showing that novel words can engage in lexical competition both with each other and with existing words in the lexicon. It seems that novel words can become full, integrated members of the lexicon and behave very much like real words, at least when the training is consistent to allow for well defined lexical representations to emerge. However, unlike the initial creation of the new lexical representation, the process of lexical integration typically requires a period of consolidation, during which sleep seems to play a

significant role. This fits in well with the notion of complementary learning systems. Having established that novel words interact with the existing lexicon, I now turn to evidence showing novel words interacting with sublexical units.

1.3.2 Perceptual learning

It is important for listeners of speech to have accurate knowledge of the phoneme categories used in their language. This is crucial for efficient word recognition, especially in discriminating between minimal word pairs, that is, words that differ only in one phoneme (e.g., *cap – gap*). However, some degree of flexibility must also be retained in order to cope with the variability in the signal. For example, the realisation of a phoneme will vary depending on factors such as the immediate phonological environment within the word, and speaker properties, such as accent. Norris, McQueen, and Cutler (2003) discussed the example of the word *total*. In British dialect both instances of /t/ in *total* are roughly identical, whereas in the American dialect the second /t/ is realised as a flap. Presumably a British listener will have to conclude that the flap produced by an American speaker is an instance of /t/, and adjust his or her phoneme category for /t/ to accommodate the new variant.

Norris et al. (2003) discussed the adjustment of phoneme categories in response to phonemic ambiguity, an event they termed perceptual learning. Participants heard words and nonwords where one of the phonemes was replaced with an ambiguous phoneme. For example, the /f/ in the Dutch word *witlof* (chicory) was replaced with the ambiguous phoneme /?fs/, half way between /f/ and /s/. After exposure to this and other similar words participants carried out a phoneme categorisation task where they categorised unambiguous and ambiguous phonemes from an /f/ - /s/ continuum. Participants were now more likely to categorise the ambiguous region as /f/, compared with a control group who had heard the same ambiguous phoneme (/?fs/) in lexical contexts supporting an /s/ interpretation. It appears that upon hearing *witlo[?fs]* the participants used lexical information to interpret the ambiguous phoneme as an /f/, and adjusted their /f/ phoneme category to accommodate the ambiguous sound. When stimuli consisted of nonwords perceptual learning did not take place, confirming the necessity of lexical information in perceptual learning. These findings have been replicated and extended in a number of recent studies (see Samuel & Kraljic, 2009, for a review), showing

that the learning effect is immediate and not dependent on sleep (Eisner & McQueen, 2006), that it is long-lasting (up to 25 min: Kraljic & Samuel, 2005; up to 12 h: Eisner & McQueen, 2006), that it is speaker-specific when using phonemes that carry information about speaker identity and not speaker specific when using phonemes that do not carry speaker information (Kraljic & Samuel, 2007; Eisner & McQueen, 2005), that the effect generalises to words not heard in training (McQueen, Cutler, & Norris, 2006), that the effect only takes place if the listener has no reason to think the ambiguity is caused by temporary idiosyncrasies of the speaker or stimulus (Kraljic, Samuel, & Brennan, 2008), and that perceptual learning on one phoneme category can generalise to another category provided that the phoneme pairs share the same primary contrast (Kraljic & Samuel, 2006).

1.3.2.1 Novel words can adjust phoneme categories

The key finding for the purposes of novel word learning in the perceptual learning literature is that nonwords cannot adjust phoneme categories. Lexical feedback is needed for this to take place. Thus one way of testing whether a novel word is capable of lexical engagement is to test if it can engage with the phoneme category level as real words do by adjusting phoneme boundaries.

Leach and Samuel (2007) did just that. They trained participants on novel words which all included either an /s/ sound (e.g., *gatersy*) or a /sh/ sound (e.g., *wikoshah*). This was followed by an exposure phase where participants heard the novel words with the /s/ or /sh/ replaced with an ambiguous phoneme half way between /s/ and /sh/. A testing phase consisting of a phoneme monitoring task was included to establish the emergence of perceptual learning effects. All phases were carried out over five consecutive days. Across five experiments the authors varied the training task. Successful and immediate perceptual learning was observed in experiments where the training attached meaning to the novel words: word-picture association and reading short passages using the novel words in context. Experiments where phoneme monitoring was used as the training task did not show reliable perceptual learning effects. Also, experiments where the training included repetition aloud failed to show the effect, irrespective of semantics. A test of explicit word form knowledge was also included in the form of a task requiring word identification in noise, as discussed earlier, which showed good form knowledge in all experiments, independent of the training method.

Compared with the lexical competition paradigm, this series of experiments included some surprising findings. Firstly, perceptual learning effects were observed only in experiments where the novel words were associated with a meaning. Recall that Dumay et al. (2004) showed no advantage for novel words trained with a semantic referent. This might be due to some of the major differences between the work of Gaskell and colleagues and the Leach and Samuel study. Leach and Samuel used richer semantic context, required participants to learn fewer words, and gave more training sessions. The second interesting finding was the fast emergence of the effect: the characteristic delay seen in lexical competition studies was absent here. The different time course of these effects and the delayed lexical competition effects was recently addressed by Davis and Gaskell (2009) who argued that such a difference is predicted by the CLS framework (see page 45 for a more detailed description of this account). According to this view the fast learning hippocampal system has a direct link to lexical phonology, explaining why novel words are able to re-tune phonemic boundaries even if they have not been lexically integrated and undergone a transfer from the hippocampal to neocortical system.

Converging evidence for this view was reported by Snoeren, Gaskell, and Di Betta (2009) who demonstrated compensation for place assimilation with newly learned spoken words. Place assimilation refers to the finding that in continuous speech certain word final consonants change according to the properties of the first consonant of the following word (e.g., the /n/ in *lean bacon* is often assimilated with the following /b/, resulting in *leam bacon*). Listeners however perceive the assimilated phoneme as an instance of the canonical form (i.e., in *leam bacon* listeners report hearing *lean bacon*), in other words they compensate for assimilation. This compensation is typically not found in assimilated nonwords, suggesting that the effect is lexically driven. In Snoeren et al. (2009) participants learned novel words ending in /t/ or /n/ (e.g., *decibot*), and were later exposed to a test phase where the assimilated form of the words were presented in a spoken sentence (e.g., the *decibo[p] behaved badly*). In a phonetic categorisation task participants were asked to judge whether the novel word in the sentence contained a target phoneme consistent with the canonical form of the word (i.e. /t/ in the current example). When the sentence provided a viable context for assimilation, participants did indeed report hearing phonemes consistent with the canonical form more often than in unviable sentence context, showing compensation. This was not the case

when tested with untrained nonwords, suggesting that there is significant top-down lexical involvement in perception of assimilation, and that only trained novel words had acquired enough of a lexical status to allow for compensation for assimilation to operate. Importantly, this was the case for words tested immediately after training and one day after training (although responses overall were faster for the words learned the day before). The lack of a consolidation effect here agrees with the perceptual learning findings, and presumably can be explained by the same mechanism in the CLS theory.

1.3.3 Semantic priming

Earlier I discussed evidence showing novel words engaging in lexical competition with each other and existing words. Lexical competition is not the only way words interact with each other, this can happen at the semantic level too. In semantic priming the presentation of a word, the prime (e.g., *nurse*), facilitates the recognition of the target (e.g., *doctor*), a semantically related word, presumably as a result of activation spreading from the prime word meaning to the related target meaning (see Neely, 1991, for a review). Semantic priming then is a good candidate for a third measure of lexical integration, provided that the novel words have been trained with a meaning.

1.3.3.1 Novel words can prime semantically related real words

As discussed in an earlier section, Perfetti et al. (2005) taught participants the meanings of previously unknown rare words, and tested word learning in a semantic decision experiment, where trials consisted of word pairs (prime – target), the first of which could be a novel word or a familiar word. Accuracy in this task was very good, showing that participants had good explicit knowledge of the meaning of the novel words. The word pairs could be semantically related or unrelated, allowing measurement of semantic priming effects. Responses were faster to the target when it was preceded by a semantically related prime, both when the prime was a familiar word and when it was a newly learned rare word. Furthermore, when a target followed an unrelated familiar or novel word, a higher amplitude N400 was detected than in the related condition. This was not the case with untrained novel words. This is important because the N400 is thought to be sensitive to the

integration of semantic information of a word with a preceding context (Kutas & Hillyard, 1980). The novel words appear to provide the required semantic context to elicit an N400 to the target. However, semantic decision is an unusual measure of semantic priming, as it requires an explicit decision to be made about the relatedness of the stimulus pair. Whether this finding would apply to more traditional measures of semantic priming will be discussed in Chapters 4 and 5.

Breitenstein et al. (2007) showed cross-modal semantic priming effects with newly learned words. Here novel words were paired with a picture over the course of five days of training in the same training paradigm as used by Breitenstein and Knecht (2002). Priming was evaluated before and after training in a primed semantic decision (living vs. nonliving) to the target pictures, with trained and untrained spoken novel words acting as primes. Comparing pre-training performance with post-training performance, responses to trials with a semantically related trained novel word prime speeded up, while no change was found in untrained novel word trials, suggesting that the trained novel words had acquired meaning. Dobel et al. (in press) replicated this finding in an experiment using magnetoencephalography (MEG) focusing on the N400m, the MEG equivalent of the N400 ERP component. Before training presenting a novel word prime with its paired picture evoked a large N400m. However, after training the N400m was present only if the prime was an untrained novel word, the N400m was attenuated for trained novel words paired with the related picture, reflecting reduced effort required in semantic processing.

Mestres-Misse, Rodriguez-Fornells, and Munte (2007) have also reported semantic priming effects with novel words in an ERP experiment similar in some respects to that of Perfetti et al. (2005). In semantic decision target words elicited a high N400 in the unrelated condition, both with real word and novel word primes. The behavioural data showed slower RTs to the related condition in both novel and real words conditions, a rather surprising finding perhaps explained by the word training method. During training the novel words were presented only three times each, in a sentence context that became semantically more constrained at each presentation. This may have led to weaker explicit knowledge of the word meanings (supported by fairly low accuracy scores in the priming test) than in the Perfetti et al. study. This explanation is supported by an earlier priming study using newly learned words. Dagenbach, Carr, and Barnhardt (1990) taught participants previously unknown rare words and their meanings. These novel words were then used as

primes in a semantic priming test using lexical decision to target words. The data showed both semantic facilitation and inhibition, depending on how well the new word meanings had been learned. Novel words which were recognised as familiar items but whose meaning was not recalled produced semantic inhibition, while novel words whose meaning was recalled produced semantic facilitation (although the latter effect was consistently observed only when instructions encouraged participants to use the prime as a predictor of target identity). The authors suggested that novel words whose meaning was poorly learned were associated with weak semantic activation, which was in danger of being obscured by activation of competing related semantic concepts. This resulted in inhibition of the competing concepts, allowing correct semantic retrieval to take place. Novel words whose meaning was well learned did not require this inhibition of competitors, allowing facilitation to take place. These data are consistent with the behavioural data of Mestres-Misse et al. (2007), where it is highly likely that novel word meanings were weakly represented due to limited training.

1.3.3.2 Novel words can inhibit semantically related real words

As just discussed, the presence of a word does not always facilitate the processing of a semantically related word. Another example of this is experiments where printed words can slow naming times of semantically related pictures when the two are presented simultaneously or with a very short stimulus onset asynchrony (SOA). One such picture-word interference (PWI) study used novel words and sought to establish whether novel words, when trained with meaning, are able to elicit PWI (Clay, Bowers, Davis, & Hanley, 2007). Presenting a novel word simultaneously with a picture slowed picture naming times relative to presenting an untrained nonword, and this effect was observed immediately after training. This general PWI effect however does not imply that the meaning of the novel word has been learned, as the same effect is observed when any word is presented with a picture, regardless of the semantic relationship between the items. Crucially, the authors also showed that presenting a novel word simultaneously with a semantically related picture slowed naming times relative to presenting an unrelated novel word. This specific form of PWI can only occur if the meaning of the novel word has been learned. Interestingly, this effect was not observed immediately after training, but only in a delayed test session which took place a week later. These data confirm that

a lexical representation can be created very quickly (the source of general PWI). However, in this experiment there was no evidence of learning the meaning of the novel words until a week after training, seemingly requiring some form of consolidation. This is surprising compared to the ERP studies where there was at least some neural evidence of immediate semantic learning. Clearly the time course of semantic learning needs more research, and this issue is one that will be studied in Chapters 4 and 5.

1.4 Factors affecting novel word learning

I have now covered a wealth of research addressing the issue of adult word learning. People seem to be able to learn the form and meaning of novel words quickly and retain this information for a long time, demonstrating the establishing of a new lexical representation. Novel words also show evidence of lexical integration: they participate in lexical competition, they can interact with a sublexical level by re-tuning phoneme categories, and the semantic activation of a novel word can spread to related meanings, as shown by priming effects. The next question I will discuss is whether some novel words are easier to learn than others. For example, some researchers have sought to establish that richer semantic information facilitates learning. Certainly the Leach and Samuel (2007) findings of only meaningful novel words enabling perceptual learning suggest meaning is not a trivial issue. There is also an interesting debate on whether phonological neighbourhood size helps or hinders the learning of new words.

1.4.1 Semantic factors

One of the earliest studies looking into the role of meaning in novel word acquisition was carried out by Whittlesea and Cantwell (1987). In their first experiment participants learned the meanings of 12 novel words. After training, the novel words were briefly presented on screen (20 ms) and the task was to report the identity of a target letter in the word. This initial experiment showed equally accurate performance with real words and the trained novel words, and significantly worse performance with untrained nonwords, suggesting the emergence of a novel lexical representation. In their second experiment Whittlesea and Cantwell compared

semantic and non-semantic training. Now equal performance was found for real words and semantically trained novel words, and significantly worse for non-semantically trained novel words (a third experiment replicated this pattern 24 hours after training, indicating persistence of facilitation). Interestingly, there was no correlation between letter-detection performance and explicit recall of the novel word meanings, again hinting at the relevance of a distinction between explicit and implicit memory traces. Balota, Ferraro, and Connor (1991) used these data (and those of Forster, 1985) to argue that meaning plays an integral part in word recognition and that an account of word learning without taking meaning into consideration is not adequate to explain the data.

Rueckl and Olds (1993) examined the effect of meaning in learning novel words using identity priming. They taught participants novel words either without meaning or with meaning. In the priming task participants saw the words briefly on screen (34 ms) and were asked to report the word. Some items were presented once during the experiment and others three times, under the assumption that repeated presentations should prime recognition. The data showed no priming effect for novel words with no meaning, and a reliable priming effect for novel words with meaning. Two following experiments manipulated meaning consistency such that either one or three different meanings were assigned to one novel word. This manipulation had no effect. Rueckl and Olds concluded that an association between a novel word and its meaning is helpful in visual word recognition, potentially due to orthographic-semantic associations, as predicted by connectionist accounts that postulate interactive connections between these levels.

Rueckl and Dror (1994) manipulated orthographic-semantic systematicity in novel words. Participants learned either a list of similar novel words with meanings from one semantic category (e.g., *durch*, *hurch*, and *kurch* paired with dog, cat, and bear) or a list of similar novel words with meanings from different categories (e.g., *durch*, *hurch*, and *kurch* paired with dog, shirt, and table). Training took place in five sessions over five weeks, with memory tasks and identification tasks carried out in each session. Cued recall tests showed faster learning for systematic novel words, although by the end of the experiment performance in both conditions was equal (and at ceiling). The identification task required participants to identify briefly presented words on screen. Furthermore, half of the novel words in the identification task had been seen just before the task in a cued recall task. Hence

half of the novel words had been primed. Overall, performance was better for systematic than non-systematic novel words. Looking at the priming conditions, the primed novel words showed no effect of semantic-orthographic systematicity, while the unprimed stimuli did. Also, only the non-systematic novel words showed priming effects. Setting the priming manipulation aside, both measures of novel word learning indicate superior performance on novel words with systematic orthographic-semantic mappings. This seems to again suggest that semantics is an important variable in word learning.

Studies looking at reading accuracy and speed have not always been successful in finding a semantics advantage. Nation, Angell, and Castles (2007) had 8- and 9-year-old children read novel words, and later tested their familiarity with the novel words in a visual 4AFC task where the foils were orthographically and phonologically similar words to the novel one. Number of exposures in training was varied (1, 2, or 4 exposures), as was semantic context. Some words were presented in the context of a story, and other in isolation. Collapsed across all conditions, performance was good in the recognition test, 48% correct one day after training, and 40% 7 days after training (chance level was 25%). Greater number of exposures at test was associated with better performance. This can be taken as another demonstration of quick learning of word form, this time in young children learning to read. The semantic manipulation however showed no statistically reliable effect, words learned in semantic context resulted in equally good performance as words learned in isolation.

A similar study by McKague, Pratt, and Johnston (2001) examined children's (6-7 year old) novel word learning in a naming task. Children were taught a number of novel words orally, either in a semantic condition (as part of an illustrated story) or in a non-semantic condition (listening to and repeating novel words). After two training session over two days, children's reading times of the novel words were measured. A free recall test was also included. All novel words were read faster than control nonwords, with the semantic manipulation having no effect. In the free recall task however semantically trained novel words were recalled more reliably than non-semantically trained items. The experiment was repeated, this time only in the non-semantic condition, with one group of children repeating the novel words aloud during training and another group learning only by listening. Again trained novel words were read more quickly and accurately than untrained

nonwords. The articulation manipulation had no effect. The authors concluded that some form of orthographic representation is formed along with a phonological representation, and that semantic or articulatory support is not necessary for this to happen.

The failure of these two studies to find an effect of semantics on reading may be due to the types of words they used. Some models of reading predict that semantics is most important when reading inconsistent words. This prediction is most strongly made by the parallel distributed model (PDP) of reading (e.g., Harm & Seidenberg, 2004). McKay, Davis, Savage, and Castles (2008) tested this prediction in a series of experiments where adult participants learned meaningful and meaningless novel words where the pronunciation was either consistent or inconsistent with real words with the same orthographic body. For example, the vowel in the novel word *trean* would be pronounced as /i/ in the consistent condition (to rhyme with *bean*), and as /e/, to rhyme with *dead*, in the inconsistent condition. Reading times and accuracies in the inconsistent condition did in fact benefit from meaning, but only when participants learned the meanings of the spoken forms of the words before being introduced to the written form. The authors argue that this was because learning the link between semantics and phonology first makes the semantic pathway the more viable option (in contrast to the pathway that bypasses semantics, linking orthography directly with phonology). No meaning effect was found in the consistent condition.

Two other findings from the McKay et al. study deserve mention. Reading times and accuracies were also measured to nonwords that were orthographic neighbours to the trained novel words. Compared to a non-neighbour baseline, learning *tren* in the consistent condition facilitated reading of nonword neighbours with the same *-ean* body. In contrast, learning *tren* in the inconsistent condition slowed the reading of nonword neighbours, presumably because there was now a competing pronunciation available for the *-ean* body. These findings are reminiscent of the lexical competition work of Gaskell and colleagues, and suggest that the novel words had been lexically integrated. Interestingly, when the same participants were tested again 6-12 months later, some of these effects were still observable. Notably, the consistent novel words were still being read faster than inconsistent novel words, and the impact on neighbours was still seen. The effect of meaning on reading time and accuracy had disappeared, although those inconsistent words whose meaning

was still explicitly recalled were also read more accurately than words whose meaning had been forgotten. Together these data suggest that in adults meaning facilitates reading, but only when looking at inconsistent novel words and when the training introduces orthography after semantics.

Another line of research relevant to the question about role of semantics comes from object recognition studies, some of which teach participants novel words as labels for novel objects. James and Gauthier (2004) trained participants to name a set of novel objects which could be attached either with a name and three semantic properties, or a name only. The names were all proper names (e.g., *John*), which means that this study is not entirely comparable with novel word learning studies, but does address the usefulness of semantics in learning new word-object pairings. The recognition accuracy data showed no difference between the condition where semantic features were assigned to each object and the condition where only a name was learned. The authors reported fMRI data which however did show a difference between the conditions: the semantic condition resulted in more activation of the left inferior frontal cortex than either the non-semantic condition or an untrained condition. This is interesting because this area has previously been linked with semantic processing, suggesting that additional semantic information was learned. In terms of behaviour on the other hand it did not provide an advantage.

Similar data were reported in an MEG study by Cornelissen, Laine, Renvall, Saarinen, Martin, and Salmelin (2004). People learned names for unfamiliar objects (ancient agricultural tools), one set included name only, a second set included name and description of the tool's function, and a third set included the functional description only. Participants were trained until they reached a criterion of 98% correct. In terms of learning there was no effect of semantics, the object names with rich semantic information (about the function of the tool) were learned equally fast as names where no additional information was provided. Unlike James and Gauthier (2004), Cornelissen et al. (2004) found no cortical differences between the conditions.

Gronholm, Rinne, Vorobyev, and Laine (2005), using the same stimuli as Cornelissen et al. (2004), also failed to find an advantage for object names which had been trained with rich semantic information. In fact, these authors found a small learning advantage for the name-only condition. A small but non-significant semantic advantage was found for patients with mild cognitive impairment (MCI)

though (Gronholm, Rinne, Vorobyev, & Laine, 2007). At the neural level, measured with positron emission tomography (PET), no differences were found in neural activation between the semantic conditions in healthy participants (Gronholm et al., 2005). The MCI patient group showed higher level of activation of a visual processing area (BA 18) with the semantically rich condition, suggesting that these patients may have created more vivid visual associations with the help of the additional semantic information (Gronholm et al., 2007).

In sum, the evidence for the role of semantics in novel word learning is mixed. Whittlesea and Cantwell (1987) as well as Rueckl and Olds (1993) provided evidence for the importance of meaning. Nation et al. (2007) and McKague et al. (2001) on the other hand showed that children learning to read novel words learned them equally well whether they knew their meaning or not, although word consistency and structure of the training regime may need to be considered (McKay et al., 2008). The object naming studies showed that attaching rich semantic information to the novel objects and their names did not tend to improve learning. In addition, Dumay et al. (2004) showed novel words engaging in lexical competition irrespective of whether participants knew what they meant or not.

Some authors have expressed surprise at the apparent lack of semantic effects (e.g., Gronholm et al., 2005). It is well known that in many memory tasks, semantic processing, or "deep" encoding, results in better memory performance. In their seminal paper Craik and Tulving (1975) asked participants to carry out tasks on real words that differed in the level of processing needed, from a shallow task (decide whether a word is printed in capital letters or not) to a task requiring deep semantic analysis (decide whether a word fits in a sentence). They found that deep learning resulted in more accurate responses in a recognition task, and also took more time to carry out than shallow learning. The latter was the case also in the object naming studies, and suggests that cognitive load of the semantic learning condition is higher than non-semantic condition, potentially causing participants to dedicate fewer resources for name learning. A very different but equally plausible account would argue that since participants knew that the semantic information was redundant for many of the tasks, they may not have performed to the best of their abilities in learning the meanings. Finally, Gronholm et al. (2005) pointed out that many participants in the non-semantic conditions reported self-generated semantic

associations in all conditions. Such associations may mask any benefit semantics might selectively offer to the conditions where words are assigned a meaning.

One of the reviewed studies suggesting that semantics is helpful (or indeed necessary) for word learning is the Leach and Samuel (2007) series of experiments discussed earlier. In this study successful lexical integration was only observed with novel words that had either been associated with a picture or given a meaning through a story context. It is not clear why this is, whether it is due to the properties of the stimuli or the measure of lexical integration (perceptual learning) will be further investigated in the next chapter. It is worth noting though that semantics did not affect the degree of word form learning even in the Leach and Samuel study. Finally, one may want to exercise caution in interpreting the object naming studies in relation to the novel word learning studies. Even in the conditions where the novel object does not have a functional description, the image of the object itself provides a semantic referent for the name. Hence the non-semantic condition is not comparable to the non-semantic conditions of the word learning studies where nothing apart from the word form was available to the learner. It is possible that in the type of learning studies discussed here any semantic information over and above a simple picture is redundant, and the degree of richness of the information is irrelevant.

1.4.2 Phonological factors

Storkel, Armbruster, and Hogan (2006) have argued, based on child word learning data, that there are two phonological properties that may play a role in novel word learning in adults: phonotactic probability and phonological neighbourhood density. Phonotactic probability refers to the frequency with which a given sound occurs at a given position in a word (i.e., positional segment frequency), and also to the frequency with which two sounds occur together (i.e., biphone frequency). Neighbourhood density refers to the number of words that differ from the target word by one phoneme. Both of these variables have been found to affect word learning in children: there is an advantage for words with high phonotactic probability and words with high neighbourhood density (Storkel, 2001, 2004). However, Storkel et al. (2006) found a different pattern in adult word learning. In a paradigm where novel words were trained embedded in a story context and

associated with a picture, adult learners produced more accurate naming responses to novel words with low phonotactic probability, and high neighbourhood density.

By analysing partially correct and completely correct responses separately, Storkel et al. were able to evaluate the influence of the two variables both at an early stage of learning (where partially complete responses were made) and at a late stage of learning (where completely correct responses were made). Only phonotactic probability affected partially correct responses, and only neighbourhood density affected completely correct responses. The advantage for phonotactically rare novel words was explained by fast triggering of word learning. High-probability novel words may be deceptively similar to existing words, thus slowing the initiation of learning. This would also explain why this variable affects only the early learning stage. Storkel et al. proposed that neighbourhood density is a critical factor in the process of integrating novel word representations in the lexicon in later stages of word learning. Hearing a high-density novel word will also activate a large number of neighbours, whose activation in turn will feed back to the phonological level. The activation at the phonological level will spread back to the appropriate lexical representations. This cycle of activation will be stronger for high-density novel words than low-density novel words, and will help to strengthen the connections of the novel words with other lexical representations and phonological representations, and in this way stabilise the new entry faster.

Jarrold and Thorn (2007) carried out another experiment where phonotactic probability and neighbourhood density were orthogonally varied. Their participants were 5-, 7-, and 9-year-old children, whose task was to learn a set of novel words, representing names of monsters seen on screen. Phonotactic probability (defined as biphone frequency only) affected learning in all age groups: words with high probability were recalled more accurately. Neighbourhood size effects on the other hand were present only in the two youngest groups, where large density had a learning advantage. The 9-year-olds showed no effect in this condition. This was interpreted as a developmental shift from a lexical association approach to word learning to a more abstractionist approach.

The discrepancy between the adult and child data is difficult to explain. In terms of phonotactic probability, children show a high-frequency advantage, and adults showed a disadvantage (in early learning only). For adults a low-probability word is a reliable indication of the word being a novel one. For children however,

many words are likely to be novel, and hence phonotactic probability is a less useful cue, potentially explaining some of the discrepancy. The disappearance of neighbourhood effects in 9-year-olds is more puzzling, considering that adults show the effect. This suggests that there may be something fundamentally different about child and adult learning, but this issue requires much more research.

Effects of phonological properties are reported in word learning studies, as discussed above, as well as studies looking at nonword recall (e.g., Roodenrys & Hinton, 2002; Thorn & Frankish, 2005). The typical finding is an advantage for nonwords with high phonotactic probability, and high neighbourhood density, although as shown by Storkel et al. (2006) the effects can also go in the opposite direction depending on the task. What is clear however is that phonological properties of the novel words matter: Storkel's finding of a neighbourhood density benefit suggests that adult learners are able to use activation in existing lexical representations to help them learn novel words. The next chapter will develop these ideas further.

1.5 Conclusions and thesis outline

Adult native language word learning is a fairly young field of research, but the review of relevant studies presented here shows that there are preliminary data throwing light on this issue at many levels of processing. We have seen that lexical representations seem to be established very quickly when a novel word is encountered repeatedly, and these new representations are durable over long time periods and without intervening exposure (e.g., Salasoo et al., 1985). People seem to acquire detailed explicit and implicit knowledge about the form and meaning of the novel words, and this knowledge is available for use and can be detected both at behavioural and neural level almost immediately.

Lexical integration however takes place with different time lags, depending on how integration is measured. Novel lexical representations seem to engage with a sublexical phonemic level immediately after training (Leach & Samuel, 2007). Integrating the novel representation with other lexical entries on the other hand requires a period of offline consolidation (possibly sleep-dependent) to occur (e.g., Dumay & Gaskell, 2007). Role of consolidation with regard to semantic representations is less clear, but there is a possibility it is required at this level too

(Clay et al., 2007), in spite of ERP evidence showing immediate semantic priming effects (e.g., Mestres-Misse et al., 2007). Why might lexical engagement at some levels take place sooner than other levels? The CLS view, when applied to word learning (see Davis & Gaskell, 2009), suggests that representations of newly learned words are initially stored in the fast learning hippocampal system. Importantly, information at this level is stored in sparse, non-overlapping representations, in order to avoid new information from interfering with existing information, or two pieces of new information interfering with each other. This nature of the representations explains why lexical competition is not seen at this early stage, as new lexical representations are yet to be integrated in the lexicon. As a result of a process of offline consolidation, the new lexical representations are integrated in the existing lexicon at the neocortical level where representations are stored in an overlapping manner, allowing lexical competition to emerge. This consolidation process, at least as far as it involves meaningless form-based representations, appears to be benefit from sleep. In this framework then any lexical process that relies on the interaction of one lexical representation with another will benefit from consolidation. Importantly though any process which does not require the new representation to have been integrated in the lexicon should be observable immediately after training.

Many further questions remain unanswered. For example, what role does semantic information play in the learning process? Is meaning necessary or useful in lexical integration? I reviewed data based on reading times and accuracy in adults and children which turn out to be inconclusive, and seem to depend on training variables and orthographic consistency. A similar contradiction was seen between the Leach and Samuel (2007) data and the lexical competition data. It may be possible to reconcile these data by considering the phonological and/or semantic properties of the novel words themselves, which play an important role in word learning. This argument is explored further in the next chapter. The roles of sleep and offline memory consolidation are critical issues as well, and looking at the neural correlates of these processes may help understand how these factors operate. While neuroimaging data are starting to emerge showing what changes in the brain during novel word consolidation (e.g., Davis et al., 2009; Breitenstein et al., 2005), we have no data on the neural events during sleep that are driving these effects. This question will be addressed in Chapter 6. Finally, whether semantic knowledge of the novel words benefits from offline consolidation in the same was as lexical

integration of word forms seems to do, is a question in need of clarification. Chapters 4 and 5 will attempt to throw some light on this issue.

Chapter 2: Meaning in word learning

2.1 Introduction

As discussed in Chapter 1, the role of meaning in novel word learning is unclear. The existing evidence is mixed: some studies have found better learning when the novel words are meaningful (e.g., Whittlesea & Cantwell, 1987), while others have found no effect (e.g., Dumay et al., 2004). In the following three experiments I will focus on the effect of semantics in word learning and lexical integration.

Recall that Leach and Samuel (2007) showed in a series of experiments that novel spoken words engage with a phoneme level only if the novel words are meaningful. This was found both in experiments where the meaning was provided by pictures of unfamiliar objects (in a word-picture matching task during training) and in experiments where the novel words were embedded in short spoken passages, followed by questions about the meaning. Experiments where training provided no meaning (phoneme monitoring) showed no reliable evidence for lexical integration.

Such findings were in stark contrast with studies looking at lexical competition. Gaskell and colleagues have in several experiments showed novel words taking part in lexical competition, as a consequence of training that did not involve meaning (phoneme monitoring in most cases). Furthermore, in an experiment which did provide meaning for the words (Dumay et al., 2004), no difference in terms of lexical integration was found between the meaningful training and purely phonological training.

Leach and Samuel (2007) provided some speculative ideas on why the data appear to be inconsistent. They pointed out that the Dumay et al. (2004) study required participants to learn a fairly large set of novel words (24, as opposed to 6 or 12 in Leach and Samuel), that there were fewer training sessions (2, as opposed to 5 in Leach and Samuel), and that the meanings of the novel words had been apparently learned quite poorly by the participants in Dumay et al. (30-44% success in free recall). Furthermore, Leach and Samuel used semantic training that provided more finely defined content (a picture or a detailed story) than Dumay et al. who presented novel words in just two different sentences (e.g., *"cathedruke is a type of vegetable"* and *"the cook served the boiled cathedruke with steak and baked potatoes"*).

Another potential explanation offered by Leach and Samuel was that lexical competition might not require as strong lexical integration as perceptual learning does. Lexical competition requires interaction between representations, while perceptual learning requires re-tuning of phonemic representations. It may be that meaning is required for this stronger form of integration to take place. Finally, one major difference between the studies was the type of stimuli used. Novel words used in the lexical competition experiments need to overlap with existing words, such as *cathedruke*, derived from *cathedral*. In the perceptual learning experiments on the other hand, no such requirement exists, and hence the stimuli used by Leach and Samuel, such as *wickoshah*, did not resemble existing words. Leach and Samuel argued that it may be easier to set up a new lexical representation for the overlapping variants, than to build a new one from scratch.

This last proposal seems highly plausible, both for the reasons stated by Leach and Samuel, but also when considering the role of semantics in learning. It is likely that hearing *cathedruke* activates the meaning of the similar-sounding real word *cathedral*. Some support for this idea comes from the data on the meaningful novel words reported by Dumay et al. (2004), who found that in a free association task a large proportion of responses to the novel words consisted of the actual base words (38-47%). It seems that if a participant experiences uncertainty about the meaning of a novel word, they choose to default to the base word meaning (proportion of responses other than the base word, its meaning, or the assigned novel meaning was 11-24%). In light of these findings, it may be the case that this type of novel word is not meaningless, even if no meaning was provided in training.

If novel words that overlap with existing words "inherit" the meaning of their familiar neighbour, we should see word-like effects also in nonwords that have been derived from existing words. A small number of studies have attempted to see if word-like nonwords in fact do activate the meaning of the real words from which they are derived. Bourassa and Besner (1998) showed that nonwords which were derived from real words by changing one letter (e.g., *deg* derived from *dog*) could prime lexical decision to semantically related real words in a visual semantic priming experiment, at least when the prime nonwords were presented only briefly (40 ms). However, the priming effect was small (less than 10 ms) and statistically significant only in a one-tailed analysis. Deacon, Dynowska, Ritter, and Grose-Fifer (2004) also used derived nonwords in a priming experiment, although here the focus of interest

was on the N400 ERP component. Nonwords were derived from real words by changing one or two letters (e.g., *contle* derived from *candle*), and together with real words were used both in prime and target positions. An attenuated N400, indicating semantic priming, was found in the condition where a derived nonword was preceded by a semantically related derived nonword (e.g., *plynt – tlee*), compared with an unprimed condition where a derived nonword was preceded by an unrelated real word (e.g., *stairs – putteffly*).

In the auditory domain Connine, Blasko and Titone (1993) showed that spoken derived nonwords could prime semantically related written real words, although they found a priming effect only when the nonwords were created by changing one phoneme by less than two phonetic features. In a second study Connine, Titone, Deelman, and Blasko (1997) showed that a more significant deviation (more than 5 features) could also result in lexical activation, when measured by phoneme monitoring latency.

The above observations support the hypothesis that word-like novel words may activate the meaning of the real words they overlap with. This is particularly the case with the spoken novel words used by Gaskell and colleagues, such as *cathedruke*, which overlap with only a small number of real words. A small cohort of highly overlapping neighbours increases the likelihood of these real word competitors becoming highly activated upon the presentation of the novel word. Furthermore, the novel word deviates from the real word competitors at a late point, extending the time during which the competitors are activated.

The question of whether this has tangible consequences in adult word learning remains to be answered. To my knowledge, only one word learning study so far has manipulated the degree to which the novel stimuli overlap with existing words. Swingley and Aslin (2007) taught 1.5-year-old children novel words that could be phonological neighbours of existing words (e.g., *tog*, neighbour of *dog*) or non-neighbours (e.g., *meb*). Knowledge of the novel words was tested by monitoring the children's eye movements in response to visual presentations of known and unknown objects, some of which had been associated with the novel words in training. When the children had to discriminate between two novel objects, one of which was named, they looked more at the named object when tested with non-neighbours, suggesting that they had learned the word form. However, when tested with neighbours, they failed to identify the named object. Another interesting finding

came from displays which included a novel object and a familiar object. If the novel object referred to a neighbour novel word, and the familiar object was the word from which the novel word was derived, children were impaired on identifying the familiar word. It seems that the novel neighbour word competed with the familiar word in these displays, a finding reminiscent of the adult lexical competition studies. The different outcomes using the different types of novel words suggest that young children at least are sensitive to the overlap between the novel word and its neighbours, and that this affects the ease with which the meanings of the novel words are acquired.

Whether these stimulus characteristics are important in adult word learning, and whether they can explain the discrepancy between the perceptual learning and lexical competition studies was examined in the series of experiments reported in this chapter. The overall hypothesis is as follows: if novel words such as *cathedruke* in the absence of trained meaning activate the existing meaning of the base word from which they are derived, then lexical integration should be observed for these words even when no meaning is trained. This should not be the case for novel words that do not overlap with existing words. Experiment 1 sought to establish the effect of orthographic overlap between novel words and real words in the degree of explicit learning of novel word forms and their meanings. Experiments 2 and 3 tested the relevance of this factor in an auditory perceptual learning experiment modelled after the experiments of Leach and Samuel (2007), but adding a manipulation of novel word type in terms of overlap with existing words.

2.2 Experiment 1

The aim of Experiment 1 was to see if the properties of the novel word form in relation to existing words have an effect on learning. Participants were taught written neighbour novel words (novel words which overlap highly with a familiar base word, such as *alcohin*) and non-neighbour novel words (variants of the neighbours manipulated to overlap to a smaller degree with the familiar base word, such as *amcohin*). [2] A meaning was also provided for each word in training.

[2] This experiment was carried out in the visual modality to allow comparison with other visual experiments presented later in this thesis, looking at consolidation effects in recall of word forms and meanings.

Neighbours were divided into two meaning conditions: neighbours with consistent meanings were items where the given meaning was closely related to the meaning of the base word (e.g., *alcohin – drink*). Neighbours with inconsistent meanings were items where the given meaning was unrelated to the base word meaning (e.g., *alcohin – flute*). Non-neighbours naturally fell in the inconsistent condition, since these novel words were designed not to evoke the meaning provided by the base word. Knowledge of the novel word forms was tested after and during training by cued recall, and knowledge of meaning by a meaning recall task. The tests were administered immediately after training in half of the participants, and the other half were tested one day later. This manipulation was included to evaluate potential consolidation effects (c.f. Dumay and Gaskell's [2007] demonstration of improving recall of word forms overnight). If the manipulation of form overlap is relevant in word learning, a difference in learning outcome should emerge between the neighbours and non-neighbours, with better recall of neighbour forms and meanings. If learners are able to access the meaning of the neighbours' base words, they should find it easier to learn the novel word meanings when the meaning is consistent rather than inconsistent with the base word meanings.

2.2.1 Method

Materials

A set of 36 novel words was selected from the stimulus set used by Tamminen & Gaskell (2008). These were all novel words which have been derived from bisyllabic and trisyllabic real base words with an early uniqueness point (e.g. *cathedruke* derived from *cathedral*). The stimulus selection for the purposes of this experiment was done largely based on the semantic properties of the base words with the aim of choosing only base words that refer to concrete nouns. After a satisfactory set had been selected, a meaning was formulated for each novel word. The meaning was selected so that it was related to the meaning of the base word from which the novel word was derived. The meaning always consisted of an object and two features of the object that made it unique, for example in the case of *cathedruke* (cathedral) the meaning was *a type of church with metal benches and no windows*. Whenever possible, the object of the novel word meaning was a superordinate of the base word meaning. This was not always possible if the superordinate had been already used for

a different word. In these cases a closely related object was chosen as the novel meaning (e.g., *clarinern* [clarinet] *is a type of flute that is made of plastic and is shrill*). The two features of the objects were selected to make it unique from any commonly encountered object of the given type. See Appendix 1 for a list of all novel words and their meanings.

The novel words used by Tamminen and Gaskell (2008) were used in the neighbour condition, as these words overlap largely with their base words both orthographically and phonologically. On average the base words were 7.4 letters long, with the neighbour novel words sharing on average the first 5 letters with the base word. Stimuli for the non-neighbour condition were generated by changing one, two, or three of the first letters of the neighbour, resulting in a pronounceable novel word which had little resemblance with the base word (Appendix 1). Hence the neighbour list and the non-neighbour list were matched in length and syllabic structure. They were also matched in summed bigram frequency to ensure that both were equally difficult to learn based on orthographic properties alone. Bigram frequency values were derived from the WordGen database (Duyck, Desmet, Verbeke, & Brysbaert, 2004).

For the purposes of the cued recall task, each novel word stimulus was associated with three cues. In two cues one letter was removed from the novel word (e.g., *cathedr_ke* and *c_thedruke*). The position of the missing letter was varied across all positions so that participants would attend equally to all parts of the novel words. These two easy cues were used during training only. The third cue was used in testing and was made more challenging by removing every other letter (e.g., *_a_h_d_u_e*). The missing letters started from the first letter of the word in half of the stimuli.

Design

The set of 36 neighbour/non-neighbour novel word pairs was pseudorandomly divided into three lists, matched in length in letters and summed bigram frequency, to be used in the three experimental conditions: neighbour – consistent meaning, neighbour – inconsistent meaning, and non-neighbour. The conditions were rotated across the three lists such that each list was used in each condition an equal number of times across participants. The inconsistent meanings condition was created by pseudorandomly shuffling the meanings across the

neighbour – inconsistent meaning and non-neighbour lists, making sure that the assigned meaning was completely unrelated to the real base word.

In order to examine possible effects of offline consolidation, half of the participants were pseudorandomly allocated to be tested immediately after training, and the other half to be tested on the following day. Participants were informed of the allocation upon arriving in the laboratory.

Procedure

Participants arrived in the laboratory on day 1 and started with a training session. Training consisted of two tasks: word-meaning matching, and cued recall. In the word-meaning matching task a trial began with a novel word presented on the computer screen, paired with a potential meaning. The task was to say whether the word-meaning pair was correct or incorrect by pressing a key on the keyboard. After a response was made, accuracy feedback was given and the correct meaning was displayed on screen. In total there were three blocks of word-meaning matching, each with two presentations of each novel word, once paired with the correct meaning and once with an incorrect meaning. On each incorrect trial a wrong meaning was randomly selected from the full list of meanings. The order of trials within blocks was randomised by the presentation software. In total then each novel word appeared six times in this task.

The three blocks of word-meaning matching were interleaved with two blocks of cued recall. In these trials one of the easy cues (e.g., *c_thedruke*) was presented on screen, and the task was to type in the complete novel word. After the response was completed, the correct word was displayed on the screen. Each novel word appeared once in each block, resulting in a total of two exposures per word in this task. Hence across both training tasks each novel word was seen eight times. This level of exposure was chosen based on pilot testing in order to reach a level of performance in cued recall and recall of meanings that was above chance but not at ceiling. This was important to make sure differences across the conditions were not obscured by floor or ceiling effects.

The test session included cued recall followed by meaning recall. The cued recall trials were identical to the training phase except that no feedback was given and the cues provided fewer letters (e.g., *_a_h_d_u_e*), making the task more challenging. Cues were presented in random order. After the response was

completed, participants were asked to rate the difficulty of recalling that particular word on a scale from 1 to 7, with 1 being very easy and 7 being very difficult. The response was made by typing the number using the keyboard.

In the meaning recall task a complete novel word was presented on screen, and participants were asked to type in the full meaning of the word. Again, once the response was completed, a rating was made regarding the difficulty of recalling the meaning. There was no time pressure in either the training or test tasks. All stimuli were presented using E-prime 1.2, which also collected the responses, running on a Windows XP PC.

Participants

30 students and staff from the University of York participated in the experiment (8 males, 5 left-handed), with a mean age of 20.0 (range = 18-31). All participants were native English speakers, reported no language-related disorders, and received course credit or cash payment.

2.2.2 Results

Data analysis

Most tasks in this experiment produced accuracy data, that is, participants made a response that was either correct or incorrect. All such data in this experiment and all following experiments in this thesis were analysed using logistic regression, which has been argued to be more appropriate and less likely to result in type I or type II errors than applying analysis of variance (ANOVA) on proportional data, even if the proportional data are arcsine corrected (Jaeger, 2008). Furthermore, I used mixed-effects models in order to simultaneously assess by-items and by-subjects effects (Baayen, Davidson, & Bates, 2008; Baayen, 2008). These analyses were carried out in R version 2.5.1 (R Development Core Team, 2007) using the *lme4* package (Bates, 2005). Whenever appropriate, I included subjects and items as random effects. Whether random slopes for the fixed effects by subjects and/or items were useful was determined for each model individually by carrying out log-likelihood ratio (LLR) tests to find the random effects structure that significantly increased the goodness of fit of the model over a model with no or fewer random slopes, within the limits of each data set (very complex random structures require

larger data sets than used in this thesis). The exact structure used in each case is reported in the text.

In building the fixed factors structure, a strategy based on model simplification was followed. A full model with the fixed factors of interest and all interactions was considered first. Interactions which included significant ($p < .05$) contrasts were kept in the model, otherwise they were dropped.[3] Elimination of non-significant interactions started with highest order interactions. Marginally significant ($p \leq .06$) effects were kept in the model if they were theoretically motivated. Also, when a variable's inclusion in the random effect structure significantly increased the fit of the model, that variable was retained as a fixed factor as well, regardless of whether it was significant (Baayen, 2008).

For each fixed effect I report the estimated coefficient (b), the t- or z-statistic associated with the coefficient, and the p-value based on the t- or z-statistic. While p-values are automatically provided by the *lme4* package for the mixed-effects version of logistic regression, this is not the case in the linear models which are used in later experiments to analyse reaction time data. In those instances Markov Chain Monte Carlo (MCMC) simulations were used to estimate p-values, using the `pvals.fnc` function provided in the *languageR* package (see Baayen, Davidson, & Bates, 2008). In this thesis I will generally not report the coefficients and their statistics for non-significant results (where $p > .05$), except if an effect is marginally significant ($p \leq .06$). Instead I shall simply state that the effect in question was non-significant. Weaker effects than that will be described in detail only if they are motivated by experimental predictions. The p-values reported in the text are uncorrected for multiple comparisons. However, Bonferroni corrected alpha levels were also calculated based on the number of contrasts examined in each model. In the text each uncorrected p-value that does not reach significance based on the corrected alpha level is marked with the symbol †. This strategy gives the reader accurate information about the significance levels while also giving information about the robustness of each contrast in the face of multiple comparisons.

[3] An alternative strategy would be to evaluate the significance of a simple effect or an interaction as a whole through model comparison using LLR tests. However, LLR tests for fixed factors have been argued to be unreliable (e.g., Bolker et al., 2009). The strategy used in this thesis follows that of Baayen (2008).

All reaction time analyses were carried out on log transformed data, in order to satisfy the assumption of normality, and to reduce the effect of outliers (Baayen, 2008). Data in figures has been re-transformed to facilitate interpretation. In all figures in this thesis error bars represent standard error (uncorrected for within-participants contrasts).

Training data

Accuracy in the word-meaning training task was analysed to see if the items in the different word conditions (Figure 1, left panel) were learned equally well during training. A mixed-effects logistic regression model with subjects and items as random factors, and word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) and training trial (six word-meaning trials) as fixed factors was fitted. No interaction between word type and trial was found, hence it was dropped from the model. Contrasts focusing on the main effect of word type showed significantly more accurate performance to neighbour-consistent words than either neighbour-inconsistent ($b = -0.740$, $z = -7.05$, $p < .001$), or non-neighbours ($b = -0.617$, $z = -5.48$, $p < .001$). No significant difference was found between neighbour-inconsistent and non-neighbour conditions.

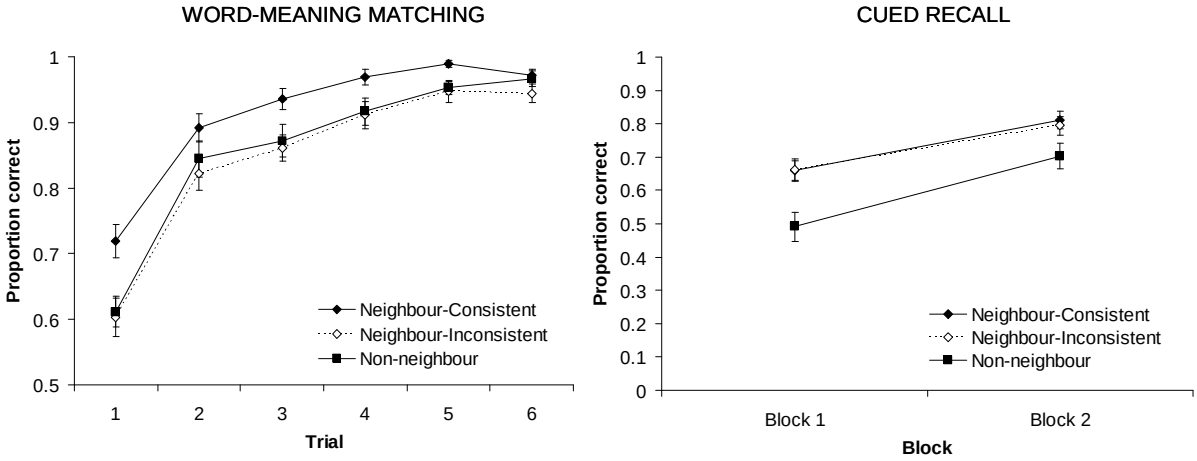


Figure 1. Accuracy rates in training tasks. Error bars represent standard error of the means.

Effect of trial was evaluated next. Accuracy improved significantly from trial 1 to trial 2 ($b = 1.253$, $z = 11.33$, $p < .001$), from trial 2 to trial 3 ($b = 0.350$, $z = 2.63$, $p = .008^{\dagger}$), from trial 3 to trial 4 ($b = 0.556$, $z = 3.51$, $p < .001$), from trial 4 to trial 5 ($b = 0.648$, $z = 3.17$, $p = .002$), but not further from trial 5 to trial 6.

Although the interaction between word type and trial failed to reach significance, visual inspection of Figure 1 suggests that performance on the three word types may have converged at trial 6. This was supported by calculating word type contrasts at each trial individually. At trial 6, the difference between neighbour-consistent and non-neighbours was not significant (b = 0.212, z = 0.48, p = .63), although the difference between neighbour-consistent and neighbour-inconsistent approached significance (b = 0.750, z = 1.88, p = .06†). There was no difference between neighbour-inconsistent and non-neighbours. Looking at the remaining five trials, the difference between consistent and inconsistent neighbours was significant at all trials (trial 1: b = 0.562, z = 3.43, p < .001, trial 2: b = 0.600, z = 2.70, p = .007†, trial 3: b = 0.887, z = 3.31, p = .001, trial 4: b = 1.156, z = 3.19, p = .001, trial 5: b = 1.628, z = 2.90, p = .004†). Similarly, the difference between consistent neighbours and non-neighbours was significant at all trials (trial 1: b = 0.523, z = 3.09, p = .002, trial 2: b = 0.433, z = 1.87, p = .06†, trial 3: b = 0.783, z = 2.86, p = .004†, trial 4: b = 1.080, z = 2.94, p = .003, trial 5: b = 1.505, z = 2.66, p = .008†). The difference between non-neighbours and inconsistent neighbours was non-significant in these trials.

Data from the two cued recall training blocks are also presented in Figure 1 (right panel). Block (block 1 and block 2) and word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) were entered as fixed factors, and subjects and items as random factors. There was no significant interaction between the two fixed factors, so the interaction was dropped. Recall accuracy for non-neighbours was significantly worse than either for consistent neighbours (b = 0.871, z = 3.47, p < .001), or for inconsistent neighbours (b = 0.817, z = 3.26, p < .001). There was no significant difference between the two neighbour conditions. Performance overall improved significantly from block 1 to block 2 (b = 1.030, z = 9.48, p < .001).

Test data

Figure 2 shows the proportion of accurately recalled novel word objects in the different word conditions, for participants who did the immediate test, and for participants who did the delayed test one day later (left y-axis). A mixed-effects logistic regression model with word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) and time of testing (immediate vs. delayed) as fixed factors and subjects and items as random factors revealed an interaction between

time of testing and word type. Hence the effect of word type was first evaluated at both test times individually. Recall rates were significantly higher for consistent neighbours compared both to inconsistent neighbours and non-neighbours: this was the case both in the immediate test ($b = -0.606$, $z = -2.26$, $p = .02^{\dagger}$, $b = -0.919$, $z = 2.90$, $p < .01^{\dagger}$) and the delayed test ($b = -1.012$, $z = -4.14$, $p < .001$. $b = -0.588$, $z = 1.97$, $p = .049^{\dagger}$). Looking at the contrast between inconsistent neighbours and non-neighbours, no significant difference was found in the immediate test or the delayed test. Looking next at the effect of time of testing, equally good performance was found at both test times in the consistent neighbour and non-neighbour conditions, but in the inconsistent neighbours condition there was a significant benefit for immediate testing ($b = 1.163$, $z = 2.45$, $p = .01^{\dagger}$).

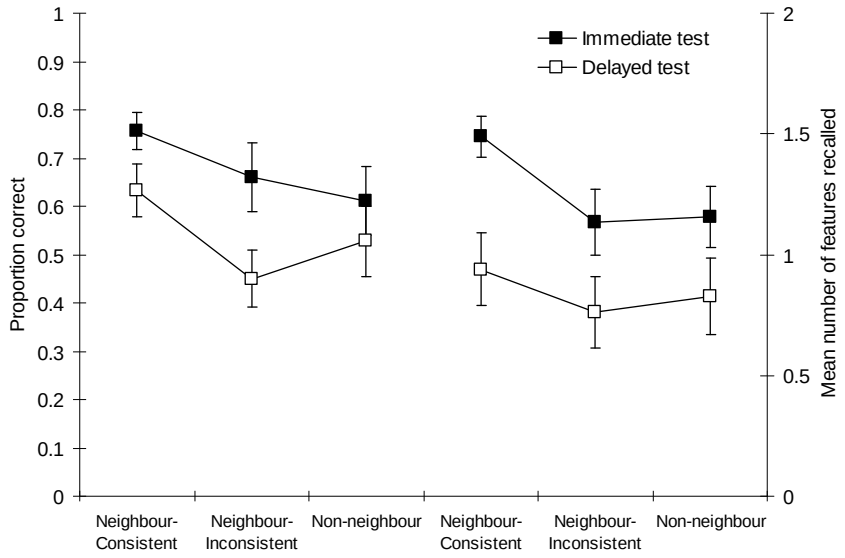


Figure 2. Recall of novel word meanings at test. Error bars represent standard error of the means.

The right y-axis in Figure 2 shows the number of features recalled in the different types of novel words and at the different test times. Logistic regression in this case is inappropriate as there are three possible outcomes: a participant can recall 0, 1, or 2 features for each word. An ANOVA could be used but is unreliable with count data. Hence ordinal logistic regression was used (Baayen, 2008; Hosmer & Lemeshow, 2000). This allows for three levels (or more) of the dependent variable and consideration of the ordinal relationship between the variables. Note that at the time of writing ordinal logistic regression is not available as a mixed model. The

regression with word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) and time of testing (immediate vs. delayed) as fixed effects showed no significant interaction contrasts, hence the interaction was dropped. The simplified model showed that more features were recalled for consistent neighbours than for inconsistent neighbours ($b = -0.598$, $z = -4.27$, $p < .001$) or for non-neighbours ($b = -0.488$, $z = 3.48$, $p < .001$). No significant difference was found between inconsistent neighbours and non-neighbours. The contrast between the two testing times was significant ($b = -0.928$, $z = -8.02$, $p < .001$), confirming that participants who were immediately tested recalled more features than participants who were tested a day after training.

Figure 3 shows the cued recall data at test. A mixed-effects logistic regression with word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) and time of testing (immediate vs. delayed) as fixed effects, and subjects and items as random effects showed no significant interactions between the fixed effects, hence the interaction was dropped. More accurate responses were made to consistent neighbours than to inconsistent neighbours ($b = -0.464$, $z = -2.53$, $p = .01$) or non-neighbours ($b = -1.207$, $z = -3.40$, $p < .001$), and to inconsistent neighbours compared to non-neighbours ($b = -0.630$, $z = 2.08$, $p = .04^{\dagger}$). The effect of time of testing did not reach significance.

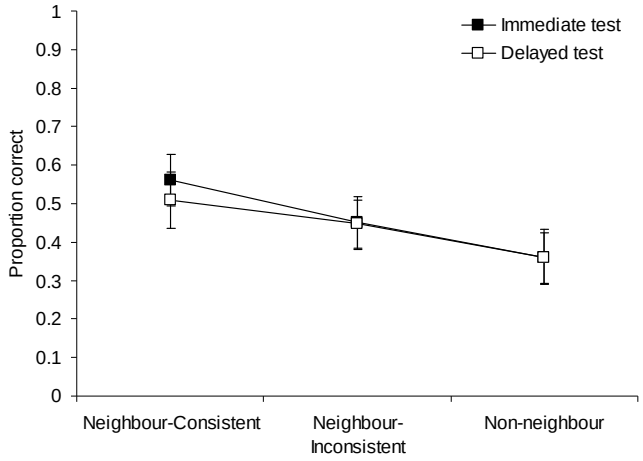


Figure 3. Accuracy of cued recall at test. Error bars represent standard error of the means.

The difficulty ratings were also analysed with ordinal logistic regression, with the rating (1-7) as the outcome variable, and word type (neighbour-consistent, neighbour-inconsistent, non-neighbour) and time of testing (immediate vs. delayed) as predictor variables. Table 1 shows the mean ratings for both tasks at immediate

and delayed test. In the meaning recall data the interaction between the predictors was non-significant and hence was dropped. Participants rated the recall of meanings significantly more difficult in the delayed test condition than in the immediate condition ($b = 0.471$, $z = 4.36$, $p < .001$). Non-neighbours and inconsistent neighbours were rated equally difficult, but the consistent neighbours were rated significantly easier to recall than inconsistent neighbours ($b = 0.674$, $z = 5.16$, $p < .001$) or non-neighbours ($b = 0.735$, $z = 5.55$, $p < .001$). In the cued recall data, the interaction between the predictor variables did not show significant effects, and was dropped. Time of testing had no significant effect, and non-neighbours and inconsistent neighbours were found equally difficult. The consistent neighbours were again rated significantly easier to recall than inconsistent neighbours ($b = 0.265$, $z = 2.01$, $p = .04^{\dagger}$) or non-neighbours ($b = 0.465$, $z = 3.55$, $p < .001$).

Table 1. Mean difficulty ratings in meaning recall and cued recall.

		Neighbour - Consistent	Neighbour - Inconsistent	Non-neighbour
Meaning recall	Immediate	3.14 (0.28)	3.93 (0.38)	4.02 (0.34)
	Delayed	3.68 (0.20)	4.39 (0.25)	4.43 (0.31)
Cued recall	Immediate	3.83 (0.28)	4.34 (0.27)	4.46 (0.25)
	Delayed	4.12 (0.20)	4.25 (0.19)	4.65 (0.17)

Note: 1 = very easy, 7 = very difficult. Standard error in parentheses.

2.2.3 Discussion

The main aim of Experiment 1 was to see if the base word meaning affects learning of novel words. This would support the hypothesis that neighbour novel words inherit the meaning of their base words. Data from training and testing supported this idea. In semantic training (word-meaning matching task), neighbours whose meaning was consistent with the base word meaning were learned faster and to a higher accuracy. In fact, the advantage was seen from the very beginning of training, with the meaning of consistent neighbours being recognised correctly 72% of the time on the first exposure. This is in contrast to significantly lower accuracies of 60% for inconsistent neighbours, and 61% for non-neighbours. This suggests that participants very quickly noticed the relationship between novel and base word meanings and were able to take advantage of it in the case of consistent neighbours.

However, one might still wonder why the two other conditions were above chance levels too. This is most parsimoniously explained by the fact that when a novel word was presented with an incorrect meaning, that meaning was randomly selected from the pool of meanings participants were being trained on. In other words, in some trials participants were able to reject an incorrect meaning based on learning in a previous trial that the same meaning was assigned to a different word.

In the cued recall training task the semantic consistency factor had no effect. Here a significant advantage was seen for neighbours over non-neighbours. This is unsurprising since in the orthographic form of neighbours can be supported by knowledge of the base word forms, and this task did not require retrieval of the word meaning.

A similar picture emerges from looking at the test data. In both recall of novel word objects and features a clear advantage was seen for consistent neighbours. No difference was found between the two conditions where the meaning-form mapping is inconsistent, and these outcomes were also reflected in subjective difficulty ratings. The cued recall data confirm the observation from training that recall of neighbour forms is superior to recall of non-neighbour forms. Perhaps surprisingly an effect of semantic manipulation was seen in this task too. The forms of consistent neighbours were recalled better than inconsistent neighbours. This was supported by participants' objective evaluation of recall difficulty. The effect is interesting since cued recall does not explicitly require access to word meaning.

There are a couple of possible mechanisms through which better form learning would be obtained in one semantic condition over the other. For example, the more difficult condition might be expected to result in more effort being allocated during learning. This however would make the opposite prediction about the outcome, with better performance in the inconsistent meaning-form condition. On a somewhat similar account, it could be that the semantic relationship between form and meaning drew participants' attention to the form during training more than in the inconsistent case. In the consistent condition the word form is a useful cue to meaning, whereas in the inconsistent condition it is a hindrance as the form provides an incorrect cue to meaning. This latter explanation makes the correct prediction about test performance.

Time of testing also had an effect in meaning recall. Significantly more features were recalled in the immediate test. The effect went in the same direction in recall of objects, although failed to reach significance. In the cued recall however there was no evidence of time of testing making a difference. The overall level of performance was lower in cued recall than in meaning recall, but not low enough to justify worries about a floor effect masking an effect of time. Thus it seems that explicit recall of novel word meanings is prone to forgetting over time, but explicit recall of novel word forms is more resistant. It is interesting to contrast this latter finding with the data from Dumay and Gaskell (2007) who in a free recall task of novel meaningless words showed a performance improvement over a night of sleep which followed shortly after training, and a non-significant decrement over the course of a day spent awake after training. There was no evidence in the current experiment of a performance enhancement over the course of a 24 hour period which presumably included sleep for all participants. This might indicate that the time between learning and the onset of sleep may be important. In the current experiment participants were trained during the day, and most likely spent several hours awake before going to sleep in the evening. This period awake may have resulted in memory decay as seen in Dumay and Gaskell (2007). Some authors have suggested that sleep may restore decay occurring during the day (e.g., Fenn, Nusbaum, & Margoliash, 2003). If this is accurate, then it is possible that participants in the current experiment experienced a decay after training, followed by a restoration of the decayed memory trace during sleep, resulting in apparently unchanged performance when tested one day after initial training. The participants who were tested immediately after training on the other hand had not experienced any decay yet. A further implication of this account would be that explicit memory for word meanings does not get restored overnight. The design of this experiment does not allow a critical evaluation of these hypotheses, but the different effect of time on semantic and form learning is interesting, and will be discussed in more detail in the following chapters. It is also worth noting that the circadian time of testing was not controlled for in this or the other experiments reported in this thesis. Young adults are at their cognitive peak in the afternoon or evening (Hasher, Goldstein & May, 2005) hence time of testing may have affected the results. However, it is likely that all circadian times (encompassing morning and afternoon) were represented and thus the effects of circadian factors would have cancelled out.

Experiment 1 showed clearly that the distinction between neighbours and non-neighbours has implications for explicit word learning, and thus is worth examining in an experiment looking at lexical integration. The experiment also presented evidence of participants being able to make use of the base word meanings in learning overlapping novel words. This lends further support to the hypothesis that neighbour novel words evoke the meanings of the base words from which they were derived, and that this may explain why lexical integration has been observed in neighbours in the absence of given meaning. This hypothesis will be put to a more stringent test in the following experiments using perceptual learning as a measure of lexical integration. Experiment 3 was a replication of the Leach and Samuel (2007) non-semantic condition, teaching participants both neighbours and non-neighbours, and evaluating perceptual learning in both groups separately. If the neighbours retain enough semantic content from the base words, it should be possible to see lexical integration in these words but not in the non-neighbours. However, before moving on to that experiment, Experiment 2 was carried out to generate and pre-test the necessary materials for perceptual learning.

2.3 Experiment 2

The primary aim of Experiment 2 was to generate a set of ambiguous phonemes and phoneme continua to be used in Experiment 3, and to make sure perceptual learning could be observed using these materials with real words. This experiment afforded an opportunity to also look at time course related issues in perceptual learning. It is now well established that the shift in phoneme boundaries caused by perceptual learning occurs immediately after exposure to the ambiguous phoneme in real word context. Furthermore, Eisner and McQueen (2006) showed that the effect does not benefit from passage of time during 12 hours, even if that time largely consists of sleep. In their experiment half of the participants were exposed to the ambiguous phoneme in the evening, and other half in the morning. Phoneme categorisation was tested immediately after exposure and 12 hours later. Hence only the evening group got to sleep prior to the second test. Both groups showed a categorisation shift immediately after training and an equally strong effect in the second test. While this shows that sleep on the first night after exposure does not significantly increase the effect, it does not completely rule out a role for offline

consolidation, which may operate over a longer time course. One purpose of Experiment 2 was to extend these findings by evaluating the effect one day and one week after initial exposure. If perceptual learning truly does not benefit from offline consolidation, then the effect should not grow stronger at this longer time course. Alternatively, while the effect may remain stable within 24 hours, it may disappear in the longer term, as no more exposure to the ambiguous phoneme is provided.

One consolidation-related aspect of perceptual learning which has not been examined yet is the time course of generalisation. Kraljic and Samuel (2006) reported an experiment where participants heard words with an ambiguous /?dt/ phoneme. As expected, participants who heard the ambiguous phoneme in a lexical context biasing /t/ were in a following phoneme categorisation task more likely to categorise sounds on a /t/ - /d/ continuum as /t/. Crucially, when these participants categorised a /p/ - /b/ continuum, they tended to respond /p/. Kraljic and Samuel argued that this is an instance of generalisation. In both contrasts the voiceless sounds (i.e., /t/ and /p/) have longer pre-release silence and longer aspiration than their voiced counterparts. Hence it appears that participants learned something not only about the phonemes in particular to which they were exposed to, but rather about the parameters of pre-release silence and aspiration used by the particular speaker. The effect was seen immediately after the exposure phase, but was not tested again later. In the current experiment the time course of perceptual learning caused by generalisation was tracked immediately after exposure, the following day, and one week later. If re-tuning as a result of generalisation is weaker than re-tuning resulting from direct exposure to the relevant phonemes, the generalised effect may decay faster. Alternatively it may benefit from passage of time if the initial effect is weak and can be consolidated. This latter view is supported by Fenn et al. (2003) who have argued that sleep-dependent consolidation is particularly helpful in generalising phonological learning to new lexical contexts (see Chapter 3 for a detailed description of this study).

In the current experiment participants first completed an exposure task where they were exposed to real words ending in /t/ and /d/, and where one of these critical phonemes was replaced with the ambiguous sound /?dt/, e.g. *awar[?dt]*. The exposure task used here was an old/new categorisation task, where participants were instructed to memorise a study list of auditorily presented words, followed by a test list which included the study words intermixed with fillers. The task was to

discriminate between the old and new items in the test list. Participants were not told about the phonemic ambiguity manipulation. The exposure phase was followed by a phoneme categorisation test on three continua: /t/ - /d/, /p/ - /b/, and /s/ - /f/. Perceptual learning effects should be observed on the /t/ - /d/ continuum. According to Kraljic and Samuel (2006) generalisation should also take place, and an effect should be found on the /p/ - /b/ continuum as well. The /s/ - /f/ continuum was included as a control that should show no effect as it does not share any phonetic features with /t/ - /d/ that might give rise to generalisation.

2.3.1 Stimulus construction and pre-test

A pre-test was carried out to construct the phoneme continua and to choose the ambiguous phoneme to be used in this and the following experiment. For this pre-test multiple tokens of the syllables /ɛt/, /ɛd/, /ɛp/, /ɛb/, /ɛf/, and /ɛs/ were recorded by a native English speaker in a sound-proof booth onto a CD, using a Sennheiser ME40 microphone, and a Marantz CDR300 CD recorder. The recordings were then copied to a PC (mono, 44 kHz sample rate, with 16 bit resolution), and edited using Adobe Audition 1.0. The /s/-/f/ continuum was created by choosing a good token of /ɛf/ and /ɛs/, and excising the frication noise from the vowel, cutting from a zero-crossing at the onset of frication, and at a zero-crossing close to the end of frication, so that both the /f/ and /s/ sounds were 221 ms long. A 21-step continuum was created by adding the amplitudes of /f/ and /s/ in different proportions, starting from a clear /f/ (100% /f/, 0% /s/) to a clear /s/ (0% /f/, 100% /s/) in increments of 5%. Because the /s/ sound was dominant, the amplitude of the original /s/ token was reduced by 5 dB before the mixing.

The /t/-/d/ continuum was created by selecting a representative token of /ɛd/ and /ɛt/, and the two tokens were aligned with regard to their respective consonant bursts. The /d/ was excised from the vowel at a zero-crossing at the onset of prevoicing. The /t/ was excised from the vowel at a point in the pre-burst silence such that the duration of pre-burst silence matched that duration of prevoicing in the /d/-token, and that the two phoneme tokens were of the same duration (169 ms) overall. A 21-step continuum was created in the same was as above. The /t/ sound was more dominating, so it was attenuated by 5 dB before mixing.

The /p/-/b/ continuum was created in the same way as the /t/-/d/ one. Here both tokens were trimmed to be 156 ms in duration, and a 21-step continuum was created. The /b/ sound was found to be more dominating than /p/, hence the amplitude of /p/ was increased by 5 dB before mixing.

After the consonant continua were created, each step on each continuum was spliced onto an /ɛ/ context, where the vowel was taken from a recording of /ɛk/, to minimise the biasing effect of coarticulatory cues in the vowel.

Eleven steps from each continuum were used in the pre-test. These ranged from one clear end to the other, in 10% increments. For each of the three continua, ten lists consisting of one token of each step were created, and the order of items within lists was randomised. These ten lists were concatenated into one long experimental list, resulting in ten repetitions of each step on a continuum. This process was carried out for each of the three continua. The order of presentation of the three continua within the pre-test session was balanced so that each continuum was presented first, second, and third an equal number of times, and that each continuum was both preceded and followed by any other continuum an equal number of times. Presentation of each continuum block started with a practice block which consisted of one presentation of each step of the continuum, in random order.

Twelve native English speakers (mean age = 19.0, 1 left-handed, 4 male) completed the pre-test in exchange for course credit or cash. E-prime 1.1 running on a Windows XP PC was used for stimulus delivery and response collection. A trial started with the presentation of an auditory token (e.g., /ɛs/), presented over headphones (Beyerdynamic DT 770), and participants were told to identify the consonant as quickly and as accurately as possible. A trial would last at most 2.6 s from the onset of the sound, or be terminated at a response. The interstimulus interval (ISI), i.e. time between a response and onset of a new sound, was 500 ms. Responses were made on a standard computer keyboard, where six keys were labelled as “T”, “D”, “S”, “F”, “P”, and “B”. For each phoneme pair, one key would be on the left side of the keyboard, and the other on the right side. Participants were asked to use their left hand to respond to the left keys, and the right hand to respond to the right keys. The allocation of responses to left and right sides was switched for half of the participants. The overall duration of the session was 15 minutes.

The responses made by participants within the trial time limit were recorded, and plotted as percentage of /b/-, /d/-, or /f/-responses, as a function of step on the

continua. These data indicated that for the /p/-/b/ continuum the most ambiguous step (categorised as /b/, /d/, or /f/ 50% of the time) was 45% /b/ mixed with 55% /p/, for the /t/-/d/ continuum it was 40% /d/ mixed with 60% /t/, and for the /s/-/f/ continuum it was 65% /f/ mixed with 35% /s/. Figure 4 shows the categorisation data for the three continua. The continuum point marked with a solid arrow was the one used as the ambiguous phoneme in this and the next experiment. Following Eisner and McQueen's (2006) method, four steps of intermediate ambiguity were also selected, in which the phoneme was classified as /b/, /d/, or /f/ for about 10%, 30%, 70% and 90% of the time. These points are marked with dashed lines in Figure 4. As that figure shows, due to the properties of the different continua it was impossible to pick steps that perfectly matched the 10, 30, 50, 70, 90 percent points, hence the step coming closest to these points was always chosen. The finely dashed points were added in the continua used in Experiment 3 to act as completely unambiguous, clear continuum end points.

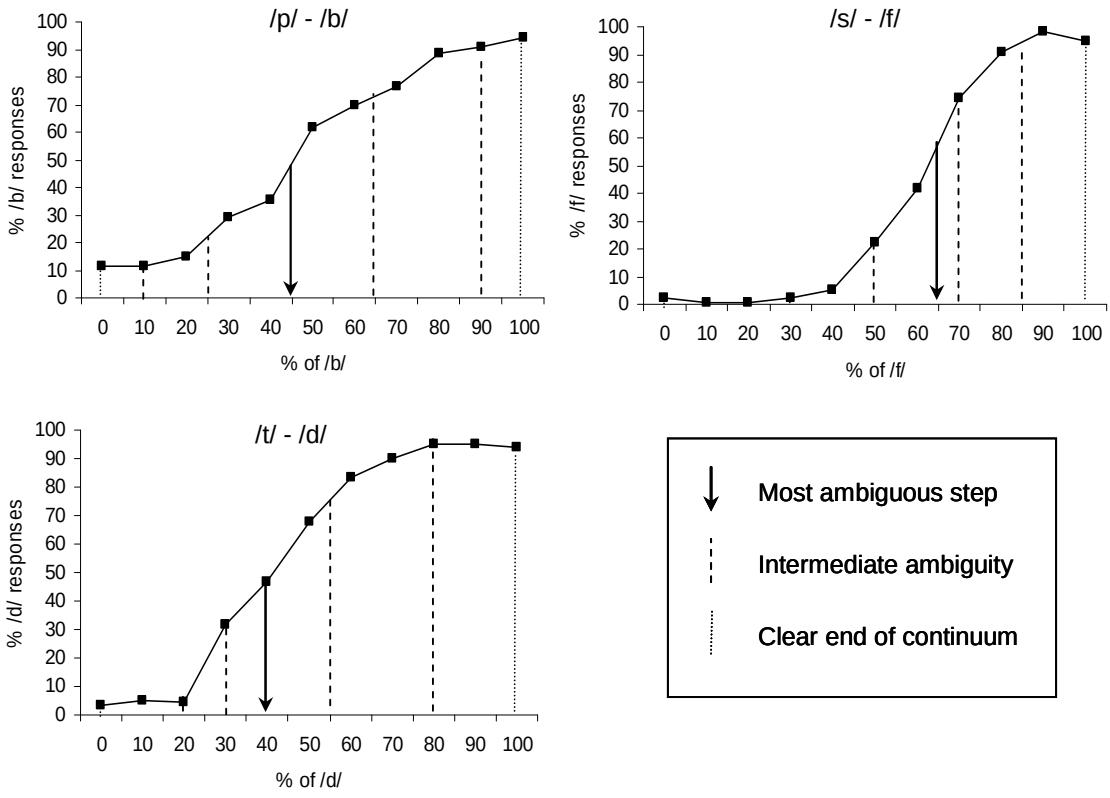


Figure 4. Categorisation functions for three continua in pre-test.

2.3.2 Method

Materials

Nineteen words ending in a /d/ sound were chosen (e.g., *award*). All words included only this one occurrence of the target phoneme. The words were bi- or trisyllabic (M = 2.1), with a mean number of phonemes of 5.0 (range 4-7). The frequencies of the words were fairly low, in order to make them better comparable to the base words used in Experiment 3, with a mean CELEX frequency (Baayen, Piepenbrock, & van Rijn, 1995) of 9 occurrences per million (range 2-30). None of the words ended in a consonant cluster. This control was added to facilitate the creation of ambiguous endings.

Nineteen words ending in a /t/ sound were also chosen (e.g. *acute*). This was again the only position where the target phoneme could occur. The properties of these words were matched with those of the /d/-ending words, in terms of frequency (M = 8.9, range 2-28), number of phonemes (M = 5.2, range 4-7), and number of syllables (M = 2.2, range 2-3). Again, none of the words ended in a consonant cluster. The two sets were also matched on stress pattern, in both sets the stress fell on the first syllable six out of nineteen times. None of the experimental words included the phonemes /b/, /p/, /f/, /s/, /v/, or /z/. See Appendix 2 for the word stimuli.

Ambiguous versions of the critical stimuli were created by replacing the final phoneme of the words with an ambiguous phoneme (/?dt/). The ambiguous phoneme was one which participants in the pre-test categorised as /t/ about 50% of the time and as /d/ in the remaining trials (see Figure 4, solid arrow).

Thirty-eight filler items were selected for the old/new categorisation task, to act as the new items. These were matched to the experimental words on frequency (M = 8.8, range 2-25), number of phonemes (M = 5.2, range 4-7), and number of syllables (M = 2.3, range 2-3). None of the fillers included the phonemes /t/, /d/, /p/, /b/, /f/, /s/, /v/, or /z/.

Design

Participants were randomly allocated to two groups. One group was exposed to the /d/-ending words in the ambiguous condition, and to the /t/-ending

unambiguous words. A second group was exposed to ambiguous /t/-ending words, and unambiguous /d/-ending words.

On day 1, all participants carried out the old/new categorisation task, the purpose of which was to expose participants to the ambiguous phoneme. Immediately after this task, they were tested on the /t/-/d/ continuum, as well as a /p/-/b/ and /f/-/s/ continua. This phoneme categorisation task was repeated on days 2 and 8.

Procedure

Old/new categorisation. Participants were presented with two lists of words auditorily, list 1 included both the 19 /d/-ending and the 19 /t/-ending words, one of these word groups ended in an ambiguous sound. Participants were asked to listen to the words in this list carefully, in anticipation of a recognition task. List 1 (study list) was followed by list 2 (test list), where the participant was asked to decide for each word whether it was an old item (a word heard in list 1) or a new item (a word not heard in list 1). The test list included all words heard in list 1, and 38 filler words. The response was made by pressing a key on a keyboard, labelled OLD or NEW. The stimuli were presented and responses collected by E-prime 1.1 running on a Windows XP PC. High-quality headphones were used stimulus delivery (Beyerdynamic DT 770).

In list 1 the words were presented with a 2000 ms ISI. Participants were not required to make any overt responses to the words in this list. The order of items was randomised for each participant by the software. In list 2 a word was presented followed by a screen asking a response to be made ("Old or New?"). Unlimited time was given to make this response. The next word was presented 1000 ms after a response was made. The order of presentation was again randomised for each participant.

At the end of this task, each participant had heard the ambiguous phoneme carried in the biasing words 19 times in list 1 and again 19 times in list 2, giving a total of 38 exposures. This is close to the original Norris et al. (2003) study where the number of exposures was 20.

Phoneme categorisation. Participants heard multiple tokens of /ɛd/ and /ɛt/ on a /t/-/d/ continuum and were asked to classify each token as a /d/ or a /t/. They also heard /p/-/b/ and /s/-/f/ continua. The /s/-/f/ continuum was always followed by the

/p/-/b/ continuum, which was followed by the /t/-/d/ continuum. This order was adopted to avoid carryover effects from categorising /t/-/d/, as in Kraljic and Samuel (2006).

All three continua consisted of five steps, ranging from a phoneme which in the pre-test was identified as /p/, /f/, or /t/ 90% of the time (step 1), to a phoneme which was identified as /b/, /s/, or /d/ 90% of the time (step 5, see Figure 4). Step 3 was the most ambiguous phoneme, and steps 2 and 4 were intermediate between the most ambiguous and the extremes.

For each phoneme pair, ten lists of the five steps in random order were concatenated into a final list of 50 trials. Hence the phoneme categorisation phase of the experiment consisted of 150 trials in total. The phonemes were presented with an ISI of 2000 ms. Participants were asked to categorise the phonemes by pressing a key on the keyboard labelled as "P" or "B" in the /p/-/b/ continuum, "F" or "S" in the /s/-/f/ continuum, and "T" or "D" in the /t/-/d/ continuum. They were asked to respond as quickly and as accurately as possible. The /f/, /b/, and /d/ responses were always made with the left hand, the opposite responses with the right hand.

Participants

Forty-two native English speaking University of York students participated in the experiment. Out of these, five failed to complete all sessions, and their data were excluded. One further participant was excluded because they had failed to respond to a large number of trials in the phoneme categorisation task (34% in the /s/-/f/ continuum). Hence the final number of participants was 36 (5 male, 4 left-handed, mean age = 19.5, range = 18-23). The participants were paid or received course credit. Two participants' data were removed from the /s/-/f/ continuum as they failed to categorise these sounds (90% – 100% of /s/ responses to all steps of the continuum on one or more days).

2.3.3 Results

Figure 5 shows the categorisation functions for the /t/ - /d/ continuum as percentage of /t/-responses. Participants who heard the ambiguous phoneme in a /t/-biasing context categorised phonemes in the ambiguous region (steps 2-4) more often as /t/ than did the participants who had heard the ambiguous phoneme in a /d/-

biasing context. This seemed to be the case in all three sessions. A mixed-effects logistic regression model including participants as a random effect, and bias (/t/-bias vs. /d/-bias), step (three middle steps of the continuum), and day of testing (days 1, 2, 8) as fixed effects was built. I focus on the ambiguous range of the continuum as there should not be an effect of bias at the end points of the continuum (same strategy was used by Leach and Samuel).

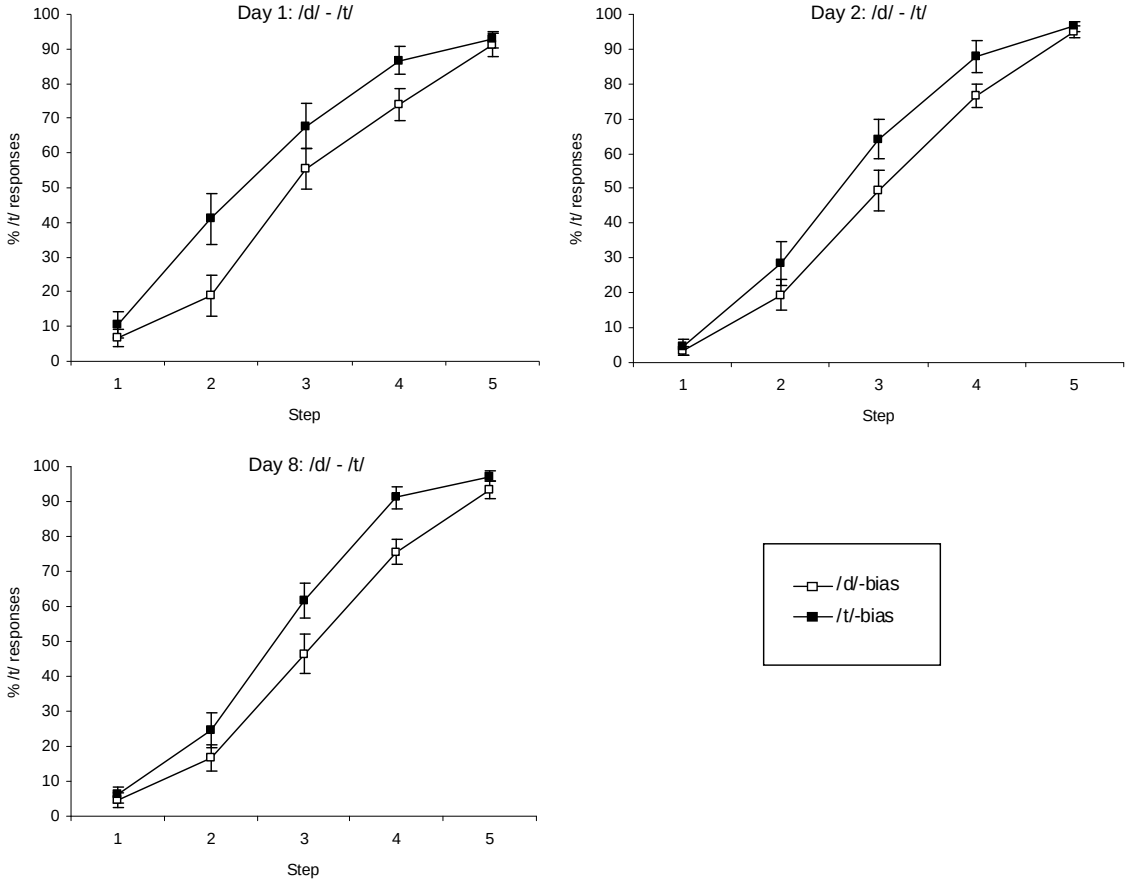


Figure 5. Phoneme categorisation data from clear /d/ to clear /t/. Error bars represent standard error of the means.

In the following analyses /d/-, /b/-, and /f/-responses were coded as "success" and /t/-, /p/-, and /s/-responses as "failure" (this information is relevant only for the purposes of interpreting the b-values, in other words I am measuring the likelihood of observing a /d/-response, positive b-values reflect an increase in this likelihood, and negative values reflect a decrease in this likelihood). LLR tests indicated that including subject-specific slopes for day and step increased the fit of the model. The full model with all main effects and interactions showed no significant interaction contrasts, so these were dropped. The most interesting effect in this analysis was that

of bias, which showed that participants who had heard the ambiguous sound in a /t/-biasing context were significantly less likely to respond /d/ than participants exposed to the same sound in a /d/-biasing context (b = -0.713, z = -2.63, p = 008[†]).

Unsurprisingly, the effect of step showed that participants were less likely to respond /d/ as the continuum moved towards /t/-like sounds (step 2 vs. step 3: b = -1.722, z = -16.30, p < .001, step 3 vs. step 4: b = -1.533, z = -10.72, p < .001). The effect of day suggested that people were more likely to respond /d/ on day 3, compared to the first day (b = 0.342, z = 2.02, p = .04[†]).

Although none of the interaction contrasts involving bias and day reached significance, it was worth evaluating the effect of bias on each day separately, to make sure the effect remains robust over time. No significant interactions were found between bias and step on any of the three days. The effect of bias was significant on all three days (day 1: b = -0.990, z = -2.43, p = .02[†], day 2: -0.711, z = -2.24, p = .03[†], day 8: b = -0.792, z = -2.70, p = .007[†]).

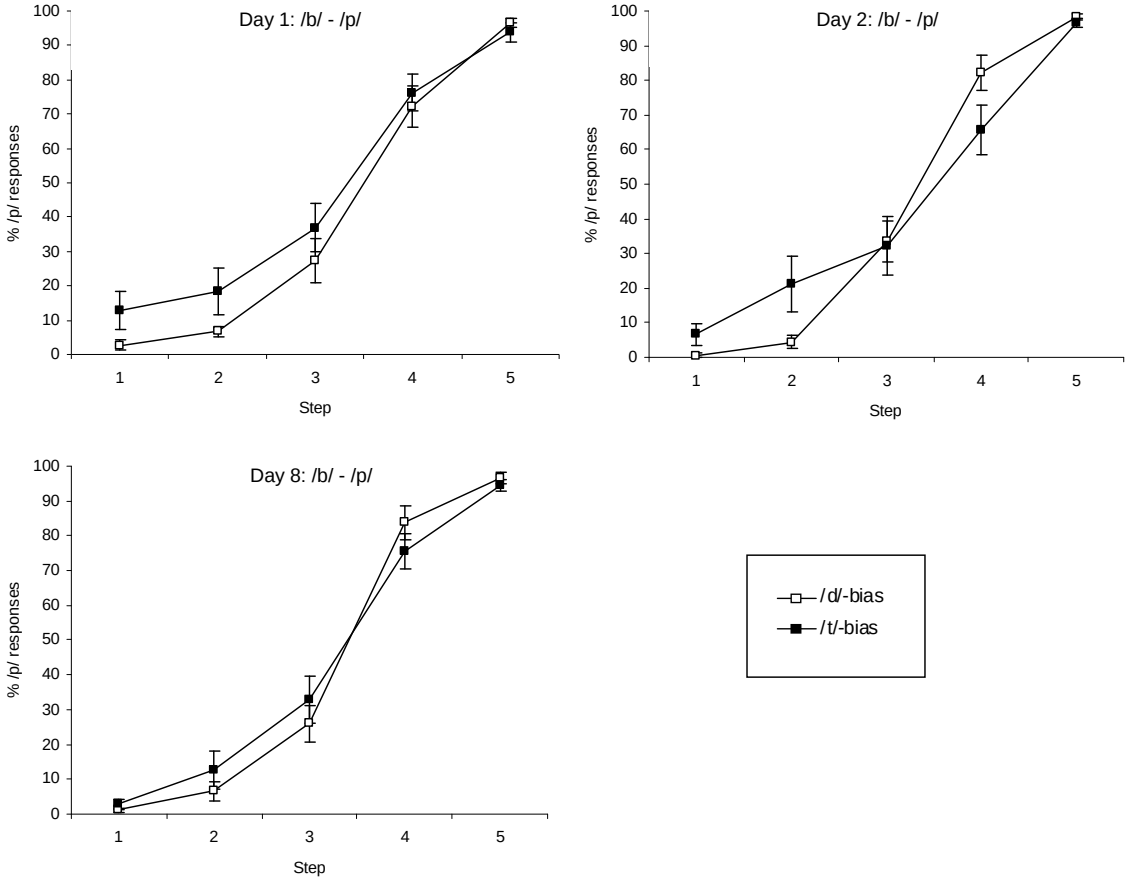


Figure 6. Phoneme categorisation data from clear /b/ to clear /p/. Error bars represent standard error of the means.

In sum, this analysis showed that, as expected, perceptual learning was found on all three days, as indicated by the effect of bias on the three days.

Responses on the /p/ - /b/ continuum were analysed next (Figure 6). LLR tests showed that subject-specific slopes for day were warranted. Here three-way interaction contrasts showed that the change in effect of bias from step 2 to step 3 was significantly larger on day 2 than day 1 ($b = 1.585$, $z = 2.14$, $p = .03^{\dagger}$) or day 3 ($b = 1.765$, $z = 2.35$, $p = .02^{\dagger}$), reflecting the reduction in the effect of bias on step 3 on day 2. The same was true of step 4 (day 1 vs. day 2: $b = 2.179$, $z = 2.85$, $p = .004$, day 2 vs. day 3: $b = 1.495$, $z = 1.90$, $p = .06^{\dagger}$), reflecting an increase in the effect of bias on day 2 at step 4, although in the unpredicted direction.

The three-way interactions showed that the pattern of data changed across the three days and steps, so each day was next also analysed separately. On day 1 the interaction between bias and step did not reach significance. The effect of bias collapsed across steps failed to reach significance, but was also examined at each step separately, and again did not reach significance at any of the ambiguous steps. Effects of step collapsed across the two bias groups confirmed that likelihood of responding /p/ increased as the steps moved towards clear /p/ (step 2 vs. step 3: $b = -1.670$, $z = -6.91$, $p < .001$, step 2 vs. step 4: $b = -4.071$, $z = -15.55$, $p < .001$, step 3 vs. step 4: $b = -2.402$, $z = -12.01$, $p < .001$). As already hinted by the three-way interaction contrasts, on day 2 bias entered into an interaction with step, whereby the effect of bias was significantly larger at step 2 compared to step 3 ($b = 1.913$, $z = 3.29$, $p = .001$), and step 4 ($b = 2.431$, $z = 3.89$, $p < .001$). The difference in bias between step 3 and 4 however did not reach significance. Looking next at the effect of bias at each step separately, the effect reached significance at step 2 ($b = -1.800$, $z = -2.42$, $p = .02^{\dagger}$) but was non-significant on the other two steps. Contrasts showed that /p/-responses collapsed across the two bias groups again increased from step 2 to step 3 ($b = -1.881$, $z = -7.21$, $p < .001$) and step 4 ($b = -4.378$, $z = -15.34$, $p < .001$), and from step 3 to step 4 ($b = -2.500$, $z = -11.90$, $p < .001$). On day 3 interaction contrasts between step and bias showed that bias at step 4 was significantly different from step 2 ($b = 1.409$, $z = 2.43$, $p = .02^{\dagger}$) and step 3 ($b = 1.107$, $z = 2.54$, $p = .01^{\dagger}$). The effect of bias when evaluated at each step separately failed to reach significance. Collapsed across the bias groups, the effect of step again confirmed that participants made more /p/-responses at the steps approaching clear /p/ (step 2 vs. step 3:

b = -1.766, z = -6.98, p < .001, step 2 vs. step 4: b = -4.654, z = -16.53, p < .001, step 3 vs. step 4: b = -2.888, z = -13.64, p < .001). This analysis then did not reveal reliable evidence for perceptual learning, as the effect of bias only reached significance on one step on day 2.

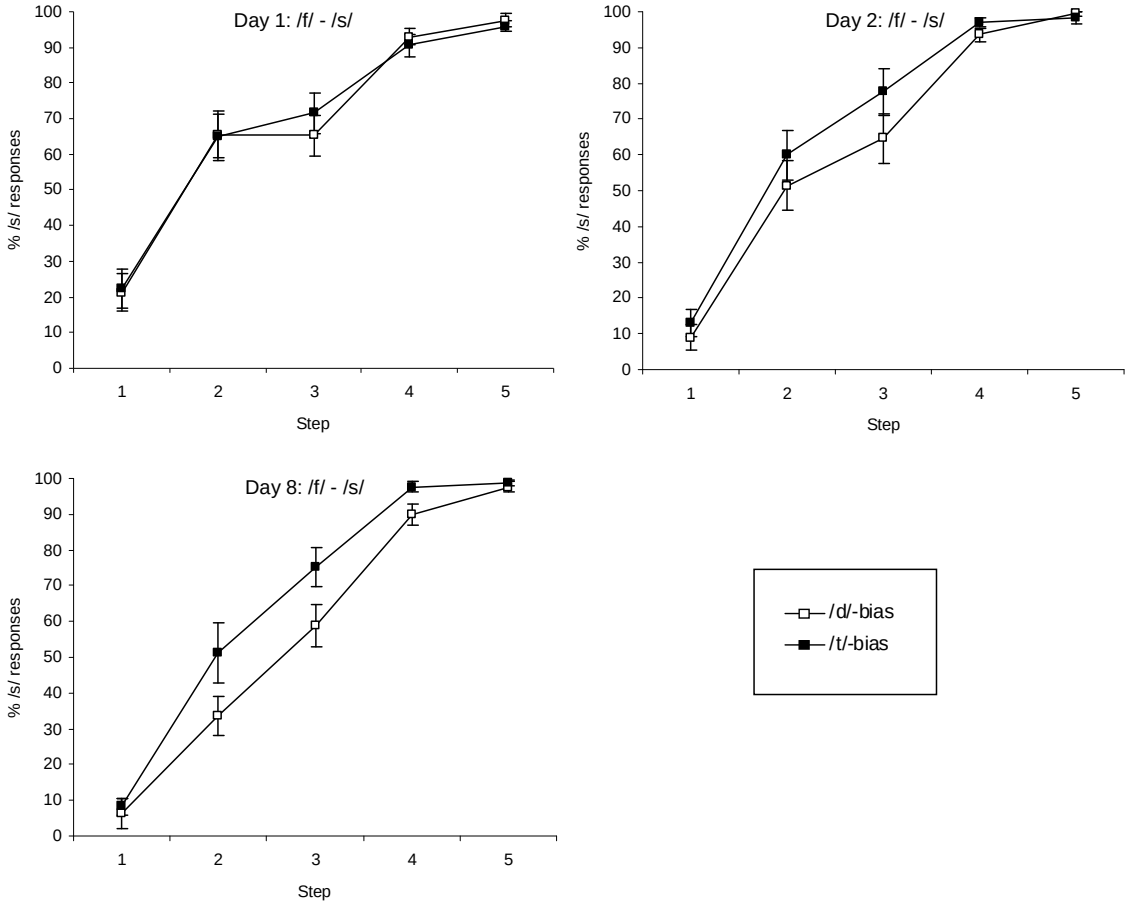


Figure 7. Phoneme categorisation data from clear /f/ to clear /s/. Error bars represent standard error of the means.

Analysis of the /s/ - /f/ continuum revealed no three-way interactions (Figure 7), so it was dropped. Bias did not enter into interaction with step, so this two-way interaction was dropped also. Bias by day interaction contrasts showed that the effect of bias increased from day 1 to day 2 (b = 0.769, z = 3.67, p < .001) and from day 1 to day 8 (b = 1.627, z = 7.57, p < .001) but not significantly from day 2 to day 8. Step and day also interacted, showing that the difference in proportion of /s/-responses between step 2 and step 3 was smaller on day 1 than it was on day 2

(b = -0.678, z = -2.67, p = .03†), and the same was true for the difference between step 2 and step 4 (b = -1.137, z = -2.96, p = .003). No further change in this regard was seen from day 2 to day 8. Due to the effect of bias interacting with day, each day was further evaluated for individually. On day 1, no interaction was found between bias and step, and the effect of bias collapsed across steps was non-significant, as it was when examined at each step individually. Proportion of /s/-responses did not increase from step 2 to step 3, but did increase from step 2 to step 4 (b = -2.139, z = -8.37, p < .001) and from step 3 to step 4 (b = -1.991, z = -7.77, p < .001). No interaction between bias and step was found on day 2 either, and here too the effect of bias collapsed across steps was non-significant (p = .12), as it was at each individual step (this is somewhat surprising in the presence of the day interaction, however the crossover on step 4 from a /d/-bias advantage to a disadvantage on day 2 increases the interaction but not the simple effect of bias. Also, the effect of bias would be marginally significant in a one-tailed analysis). Now /s/-responses increased from step 2 to step 3 (b = -0.909, z = -4.83, p < .001), from step 2 to step 4 (b = -3.471, z = -10.85, p < .001) and from step 3 to step 4 (b = -2.567, z = -8.09, p < .001). On day 8 no interactions were found, but the effect of bias collapsed across the three steps now reached significance (b = -1.132, z = -2.50, p = .01†). As suggested by the lack of an interaction between step and bias, the effect of bias was significant at all ambiguous steps (step 2: b = -0.965, z = -1.98, p = .048†, step 3: b = -1.123, z = -2.26, p = .02†, step 4: b = -1.915, z = -2.58, p = .01†). As on day 2, /s/-responses again increased from step 2 to step 3 (b = -1.333, z = -7.27, p < .001), from step 2 to step 4 (b = -3.695, z = -12.93, p < .001), and from step 3 to step 4 (b = -2.349, z = -8.44, p < .001). To summarise, the above analysis revealed an unexpected perceptual learning effect on day 8 only.

2.3.4 Discussion

The primary aim of Experiment 2 was to replicate the basic perceptual learning effect with the ambiguous phonemes created for the current set of experiments. This was achieved: people who in the exposure task heard /?dt/ in a lexical context supporting a /d/ interpretation (e.g., *award*) were more likely to categorise ambiguous sounds on a /t/-/d/ continuum as /d/ than people who heard the

same /?dt/ in a context supporting /t/ (e.g., *acute*). The effect remained reliable one day, and one week after the exposure task.

Kraljic and Samuel (2006) showed that a bias created with /?dt/ extends to a /p/-/b/ continuum. People who were biased towards /t/ were also biased towards responding /p/ on a /p/-/b/ continuum. The data in Experiment 2 did not unequivocally replicate this finding. A numerical trend for a bias in the /p/-/b/ continuum was observable on the first ambiguous step (step 2, see Figure 5), but it reached statistical significance on day 2 only. The weakness of this effect might be explained by the choice of steps for this continuum. As Figure 4 shows, the pre-test data for this continuum were less clear than for the other continua. The slope of the categorisation function is shallower and shows less of a categorical shift. This may reflect noise in the data and a lack of agreement in the point where participants shifted from responding /p/ to /b/. Hence it is possible that the sounds chosen to make up the ambiguous region of the continuum in the main experiment were not as ambiguous as the same steps in the other continua, and this was why the trend was so weak.

The emergence of a bias effect on day 8 in the /f/-/s/ continuum was surprising. Here participants who heard the /?dt/ in /t/-biased condition categorised the ambiguous /?fs/ as /s/ more often than the /d/-biased group. This continuum was included as a control condition where no generalisation was expected to occur, as /s/ and /f/ do not share the place or manner of articulation with /t/ and /d/, and both are voiceless, so the voicing contrast should not enable generalisation either. One possibility is that the ambiguous /?dt/ had some properties that made it resemble /s/ more than /f/. For example, the aspiration from the /t/ is made unusually prominent by the blending of the original /t/ and /d/ tokens, and this aspiration may resemble a token of /s/. Repeated exposure to an /s/-like sound may have resulted in increased bias towards /s/. Figure 7 indicates that on day 1 all participants appeared to make /s/-responses more than would be expected by chance (between 60% and 70% at steps 2 and 3). This bias seems to decrease in the following days, but it is unclear why it would decrease more for the /d/-biased participants. Possibly the repeated act categorising the ambiguous /?dt/ as /d/ leads these participants to ignore the superfluous aspiration inherent in the ambiguous sound. This issue would require more experimental work with careful phonetic manipulation of the /?dt/ to be solved.

In Experiment 2 the perceptual learning effect was observed immediately after exposure, as was the case with the data reported by Eisner and McQueen (2005). They also included a sleep manipulation where one group of participants was tested after sleep and another after the same time of wakefulness. This manipulation had no reliable effect, although perceptual learning was numerically stronger in the sleep group. In the current experiment a significant perceptual learning effect was found immediately after exposure to the ambiguous stimuli, and one day later. The test was repeated one week after exposure, and the initially observed perceptual learning effect was still significant. There was no evidence of the effect either increasing or decreasing over time. This suggests that the effect is highly resistant to decay. Eisner and McQueen argued that the fast re-tuning of phoneme representations is an adaptation to the need to quickly adjust to new talker idiosyncrasies and should be effortless also because the perceptual system is not required to learn anything new, rather simply to adjust the processing of a certain phoneme. This adjustment is optimal if it does not depend on consolidation and is durable over long time periods. The current data showed it is indeed durable over several days but there was no evidence of the effect growing stronger over a week, further confirming that the perceptual learning effect does not benefit from offline consolidation in the short or the long term. It might be argued that since the only hint of a generalised phoneme boundary shift on the /p/-/t/ continuum was only observed on day 2 and disappeared by the last test session the generalised effect may in fact benefit from consolidation initially but not remain statistically robust in the absence of further exposure. This would suggest that the initial hippocampal memory of the ambiguous phoneme is specific to the phoneme heard in training, and can only be generalised once a neocortical representation is generated after consolidation. However, the fact that the effect of bias was seen in only one step of the continuum on day 2 indicates that the effect may have been spurious. Future research should shed more light on this issue.

2.4 Experiment 3

As outlined earlier, Experiment 3 was designed to evaluate lexical integration in both neighbour novel words, and non-neighbour novel words by means of perceptual learning. A non-semantic training task was used (phoneme monitoring),

and perceptual learning effects were tested immediately after the first training session, and again one day later after a second training session. The second training and testing session was included as Leach and Samuel's data suggested that the effect does not always reach statistical significance after only one day of training. The Leach and Samuel (2007) findings predict no perceptual learning in this experiment since no meaning was given during training. However, if neighbour novel words evoke the meaning of their overlapping base words, perceptual learning may be observed for these novel words only.

2.4.1 Method

Materials

Twelve neighbour novel words were chosen from the list of items used by Tamminen and Gaskell (2008). The 12 novel words were all bi- or trisyllabic (M = 2.6), with a length of 6.3 phonemes on average (range 5-7), and base word mean frequency of 3.8 per million (range 2-9). Two versions of each neighbour novel word were created, ending in a /t/ or /d/ (e.g. *methanat, methanad,* from base word *methanol*).

Twelve non-neighbour novel words were created by changing the first three phonemes of the neighbour novel words described above (e.g. *piranat* derived from *methanat*). These were pronounceable words, but they deviated from real words at an earlier point than the neighbour novel words. The neighbours deviated from real words (their respective base words) on average at the fifth phoneme, whereas the newly created novel words deviated at the third phoneme, hence making them more dissimilar to any real words. Again, two versions of these novel words were created, ending in /t/ or /d/, as above. These materials are presented in Appendix 3.

Real words ending in /t/ and /d/ were also chosen, six of each. These were used as fillers in list 1 of the exposure task. These filler words were matched with the novel words in terms of number of syllables (M = 2.3 and 2.5 for /t/ and /d/ respectively) and phonemes (M = 6.5 and 6.5 for /t/ and /d/ respectively). They included no instances of the critical phonemes apart from the final position.

Finally, 12 more filler words were chosen to be used in list 2 as distracters. These were matched to the novel words in number of syllables (M = 2.8) and phonemes (M = 6.3). The frequency of list 2 filler words was overall similar to list 1

filler words (M = 4.7 for list 2, M = 21.7 for list 1, but this is inflated by one item, M = 8.9 without it). Twelve nonwords were derived from bi- and trisyllabic real words to act as nonword fillers in list 2, so as not to make the novel words stand out. These too were matched in number of syllables (M = 2.7) and phonemes (M = 6.3) to the other materials. None of the list 2 filler materials included the critical phonemes.

The stimuli were recorded by the same speaker using the same equipment as in Experiment 2. An ambiguous version of each novel word was created by replacing the final phoneme of the words with an ambiguous phoneme. Same token of /?dt/ used in Experiment 2 was also used here.

Design

Novel word type was manipulated between-participants, with half of the participants learning neighbours and the other half non-neighbours. Half of the participants within word type condition were exposed to the /d/-ending version of the novel words, and the other half to the /t/-ending version. Participants were randomly allocated into these groups.

In Experiment 2 the critical test of perceptual learning was in comparing participants who had been biased to respond /t/ by the exposure task with another group who were biased to respond /d/. While this is the most commonly used design in the perceptual learning literature, it has two weaknesses. Firstly, it provides no pre-exposure baseline measure, hence it is theoretically possible, although unlikely, that the difference between participant groups was accidental. Secondly, a between-participants manipulation is statistically less powerful than a within-participants manipulation where participant-specific variation is brought under control. With these considerations in mind, Experiment 3 used a design which allowed a within-participants comparison. All participants completed a baseline phoneme categorisation test prior to novel word training and exposure to the ambiguous stimuli. The phoneme categorisation task was repeated at the end of the exposure phase. One possible concern here is that exposure to the categorisation task might bias participants' perception of the ambiguous phoneme in a similar way as hypothesised in connection with Experiment 2 (p. 77). However, since different groups in the present experiment were lexically biased towards /t/ or /d/, this potential confound can be excluded if an effect of bias is seen in both groups.

The experiment was carried out over two consecutive days. On day 1, participants completed the phoneme categorisation twice (baseline and first post-learning test), and carried out the first training session and the exposure task. They also completed a cued recall task to evaluate explicit learning of the novel word forms. The second day was identical except that only one phoneme categorisation task was included.

Procedure

Phoneme categorisation. Day 1 started with a phoneme categorisation task, which provided the baseline against which subsequent performance after novel word training and exposure to the ambiguous phoneme could be compared. This task was identical to the same task in Experiment 2 apart from two modifications. Firstly, participants were only tested on one continuum: /t/-/d/. Secondly, two steps were added to the continuum, they were clear tokens of the phonemes added to the continuum endpoints (the endpoints in Figure 4). The reason why these clear tokens were added was to make it easier to identify participants who were responding abnormally. Unequivocal responses ought to be expected at these endpoints. If a participant deviates to a great degree from this baseline, then the decision to exclude such a participant would be well founded. [4] The critical value chosen was 30%: if a participant categorised one of the clear phonemes incorrectly 30% or more of the time, their data were excluded. In all other respects the procedure was identical to Experiment 2.

Novel word training. Participants were familiarised with 12 novel words in a phoneme monitoring task. On each trial, a capital letter indicating the target phoneme appeared in the middle of the screen for 1500 ms, followed by auditory presentation of the novel word. The letter remained on the screen until a response was made. Participants were asked to indicate whether the target was present in the word by pressing one of two keys on the keyboard, labelled YES or NO. Each novel word was repeated 24 times. The target phoneme was never the final phoneme, and consisted mostly of consonants. The possible targets included p, l, k, n, m, h, r, i, b, th, g, a, o, e. The target phoneme was absent in 50% of the trials, and present in 50%.

[4] A similar strategy was used by Leach and Samuel (A. G. Samuel, personal communication, September 16, 2007).

Each target phoneme occurred an equal number of times in the absent and present conditions. The order of trials was randomised for each participant by the software. This task took about 30 minutes to complete, and participants were given a break half way through the trials.

Old/new categorisation. This task was used to expose participants to the ambiguous /?dt/ in the novel word context. As in Experiment 2, participants again heard two lists of words. List 1 included 12 novel words and 12 real words. The final phoneme of each novel word was replaced by the ambiguous phoneme. If the participant was exposed to the ambiguous phoneme in /d/-biased lexical context, the filler real words consisted of 12 clear /t/-ending words. Those participants who were being exposed to /t/-biased contexts heard /d/-ending filler words. The task was to listen to the words carefully, and try to remember them in the forthcoming recognition task. List 2 then followed, and participants were required to decide for each item in this list whether it had been heard in list 1 or not. This list included all items from list 1, and 24 fillers which had never been heard before (12 words, 12 nonwords). Responses were made by pressing a key on the keyboard, labelled YES or NO. The order of items in both lists was randomised for each participant by the software. By the end of the task, the participant had heard the ambiguous phoneme 24 times: 12 times in list 1, and 12 times in list 2.

Cued recall. The day 1 session ended with a cued recall task. In this task a cue was played through the headphones, and the participant was asked to recall the novel word to which the cue referred, and to say the word aloud. The cue consisted of the first two phonemes of a novel word, recorded by a female native English speaker, naïve to the experiment and to the complete novel words. One advantage of using a new speaker for the cues was that it eliminated the chance of doing the task by referring to episodic memory traces of the novel words spoken by the original male speaker. Changing the speaker increases the likelihood of abstract novel lexical representations being probed. After the presentation of the cue, the participant had the option of either listening to the cue again, as many times as he or she desired, or making a response. Once a response was made, a key press initiated a new trial. The cues were presented for recall in random order. Vocal responses were directly recorded onto a minidisc (Sony MZ-N710) through a microphone built into the set of headphones (Beyerdynamic DT 294). Apart from this, the same hardware and software were used for stimulus presentation and recording as in Experiment 2.

The participants were asked to return on the following day, and carried out the same tasks, in the same order, as on day 1, except for the first phoneme categorisation task (baseline) which was omitted on day 2.

Participants

Forty-four native English speaking University of York students participated in the experiment. Out of these, six failed to attend the second session, and their data were excluded. Two further participants were excluded, one because they categorised a clear token on the phoneme continuum inconsistently, and the other because they confused the keys in the phoneme categorisation task. The final number of participants was 36 (9 male, 3 left-handed, mean age = 20.0, range = 18-42). The participants were paid or received course credit.

2.4.2 Results

2.4.2.1 Neighbour novel words

Figure 8 shows the phoneme categorisation functions for /t/- and /d/-biased participants who learned neighbour novel words. In both cases there appears to be a boundary shift in the direction predicted by novel word lexical bias, and this shift seems to be present in the three middle steps which constitute the ambiguous region. Logistic regression was used to analyse the data. For the purposes of the regression all responses that were consistent with the lexical bias were coded as "successes" (1) and responses that were not consistent with the bias were coded as "failures" (0). The main focus of the analysis was on effect of time of testing. More bias-consistent

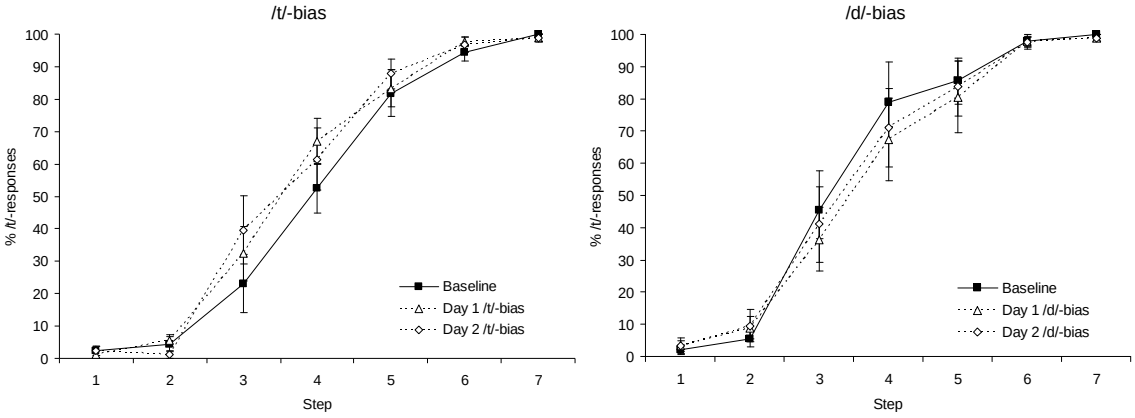


Figure 8. Phoneme categorisation data for participants who learned neighbour novel words. Error bars represent standard error of the means.

responses should be observed on day 1 and day 2 tests compared to the pre-exposure baseline. As in Experiment 2, only the ambiguous region (steps 3-5) was used in the analysis.

A mixed-effects logistic regression model with subject as random factor and testpoint (baseline, day 1 test, day 2 test), step (steps 3, 4, 5) and bias (/t/-bias and /d/-bias) as fixed factors was fitted. Subject-specific slopes for day and step improved the fit of the model. No three-way interactions were found, hence they were dropped. A model including two-way interactions showed that only the step by bias interaction was significant, and was retained in the model. In this simplified model testpoint showed a significant effect: participants were more likely to make bias-consistent responses in the day 1 test compared to baseline ($b = 0.684$, $z = 2.53$, $p = .01^{\dagger}$). The effect in the day 2 test approached significance ($b = 0.563$, $z = 1.79$, $p = .07^{\dagger}$). The interaction between step and bias simply reflected the fact that in the /t/-biased group the likelihood of making bias-consistent responses (i.e., /t/-responses) increased from step 3 to step 4 ($b = 1.418$, $z = 7.09$, $p < .001$), from step 3 to step 5 ($b = 2.801$, $z = 11.94$, $p < .001$), and from step 4 to step 5 ($b = 1.388$, $z = 6.26$, $p < .001$), while in the /d/-biased group the likelihood of making a bias-consistent response (i.e., /d/-response) decreased from step 3 to step 4 ($b = -2.212$, $z = -8.44$, $p < .001$), from step 3 to step 5 ($b = -3.364$, $z = -10.77$, $p < .001$) and from step 4 to step 5 ($b = -1.151$, $z = -4.04$, $p < .001$). Note that this change in the effect of step is because the test continuum was the same for both bias groups, but whether a response was bias-consistent or not depended on assignment to bias group. So while both groups made only a small percentage of /t/-responses and consequently a large percentage of /d/-responses at step 3 (Figure 8), the /t/-responses were bias-consistent only for the /t/-group. For the /d/-group it was the /d/-responses which were bias-consistent.

Although there was no significant interaction between the effect of testpoint and bias group, data for both bias groups were analysed separately to see if the weak day 2 effect is stronger in one bias group than the other. Visual inspection of Figure 8 suggests it is smaller in the /d/-bias group, and stronger in the /t/-bias group. For the /t/-biased group, collapsed across the three steps, significantly more bias consistent responses were made at the day 1 testpoint compared to baseline ($b = 0.497$, $z = 2.37$, $p = .02^{\dagger}$), and at the day 2 testpoint ($b = 0.636$, $z = 3.02$,

$p = .003^{\dagger}$) compared to baseline. In the /d/-biased group significantly more bias-consistent responses compared to baseline were made on day 1 ($b = 0.723$, $z = 2.86$, $p = .004^{\dagger}$), but the effect was non-significant on day 2 ($b = 0.355$, $z = 1.40$, $p = .16$).

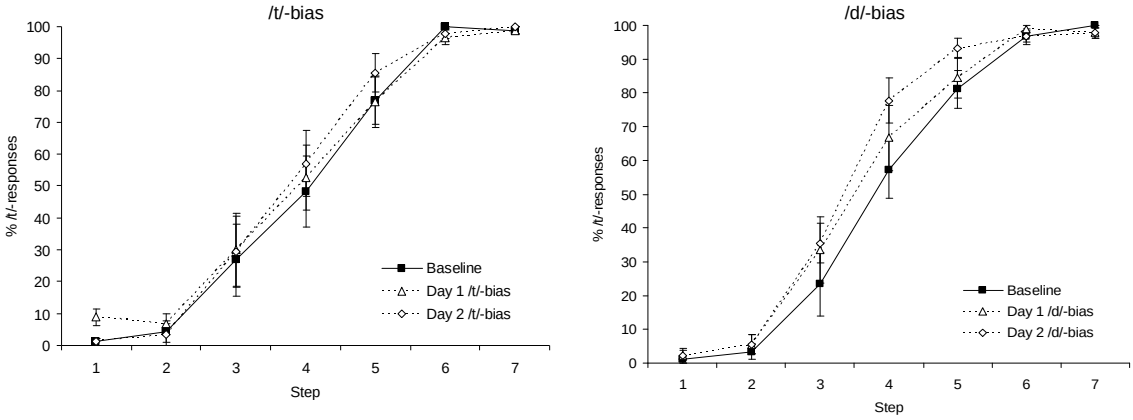


Figure 9. Phoneme categorisation data for participants who learned non-neighbour novel words. Error bars represent standard error of the means.

2.4.2.2 Non-neighbour novel words

Data for participants who learned non-neighbours are displayed in Figure 9, and were analysed in the same way. Again, subjects were included in random factors, and testpoint (baseline, day 1 test, day 2 test), step (steps 3, 4, 5) and bias (/t/-bias and /d/-bias) as fixed factors. Subject-specific slopes for testpoint improved the fit of the model. No three-way interactions were significant. Of the two-way interactions, contrasts involving a step by bias interaction were again significant. As before, this showed that in the /t/-biased group bias-consistent /t/-responses increased from step 3 to step 4 ($b = 1.438$, $z = 6.29$, $p < .001$), from step 3 to step 5 ($b = 3.015$, $z = 11.82$, $p < .001$) and from step 4 to step 5 ($b = 1.581$, $z = 7.09$, $p < .001$). In contrast, in the /d/-biased group, bias-consistent /d/-responses decreased from step 3 to step 4 ($b = -1.952$, $z = -8.94$, $p < .001$), from step 3 to step 5 ($b = -3.290$, $z = -12.60$, $p < .001$), and from step 4 to step 5 ($b = -1.399$, $z = -5.59$, $p < .001$).

Here also testpoint entered into an interaction with bias, showing that /d/-biased participants showed a larger shift from baseline to day 2 test than /t/-biased participants ($b = 1.433$, $z = 4.50$, $p < .001$). The same contrast for day 1 approached significance ($b = 0.601$, $z = 1.93$, $p = .054^{\dagger}$). Note however that the larger shift in /d/-biased participants is not in the direction predicted by bias, and hence does not demonstrate perceptual learning in this group. No interaction was found between day

and step. Since testpoint interacted with bias group, it is important to see if testpoint reaches significance in either bias group separately. Collapsed across steps 3-5, in the /t/-biased group there was no significant difference between baseline and day 1 testpoints. However, at day 2 the effect reached significance (b = 0.446, z = 2.00, p = .046†). The /d/-biased group showed a significant difference between baseline and day 1 (b = -0.452, z = -2.06, p = .04†) and between baseline and day 2 (b = -0.988, z = -4.35, p < .001). However, as already noted, this shift represented an increased likelihood of making a /t/-response, which was not consistent with the lexical bias.

Figure 10 summarises the difference in perceptual learning between neighbour and non-neighbour conditions. The proportion of bias-consistent responses was calculated for the baseline and the two post-exposure tests. The figure shows the difference between baseline and tests collapsed across the three ambiguous steps and across bias groups. It highlights the observation that hearing the ambiguous phoneme in the neighbour novel words initiates a shift in phoneme categorisation boundaries (black bars), while hearing the same ambiguous phoneme in non-neighbours results in no significant shift (white bars).

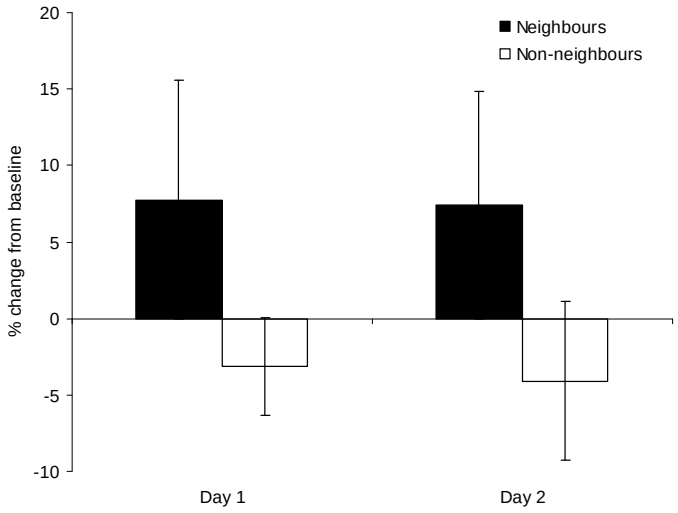


Figure 10. Categorisation change from baseline in the ambiguous range on the two testing days. Error bars represent standard error of the means.

2.4.2.3 Cued recall

Figure 11 shows the accuracy rates in cued recall. Responses that deviated from the novel word by one or more phonemes, as well as failure to make a

response, were scored as incorrect. One participant's data in the non-neighbour condition were excluded as the participant misunderstood the task and produced no data suitable for scoring. One further non-neighbour participant's day 2 data were lost due to equipment malfunction. A mixed-effects logistic regression with subjects and items as random effects, and testpoint (day 1 test, day 2 test) and word type (neighbours and non-neighbours) as fixed factors showed no interaction between the two variables. The effect of testpoint was significant, reflecting increasing accuracy on day 2 ($b = 1.788$, $z = 9.43$, $p < .001$) compared to day 1. No significant effect of word type was found however, suggesting that both word types were learned explicitly equally well.

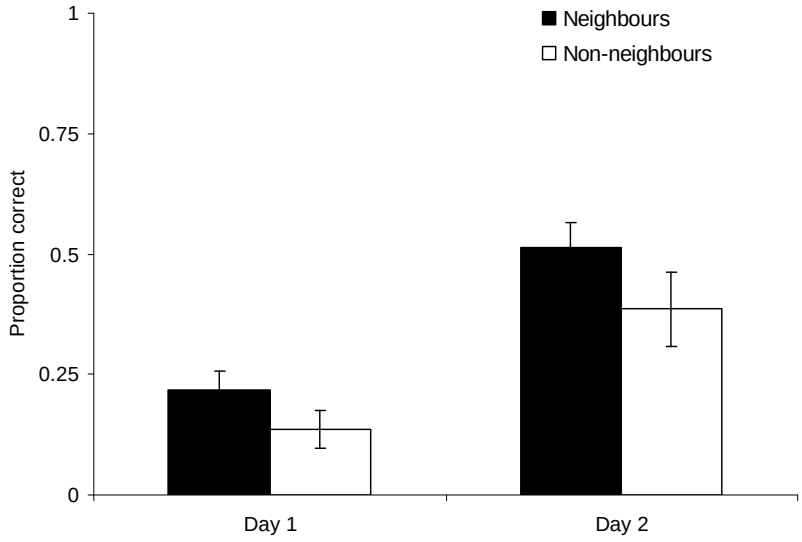


Figure 11. Accuracy rates in the cued recall test. Error bars represent standard error of the means.

2.4.3 Discussion

Experiment 3 was consistent with the findings of Leach and Samuel (2007), in that no evidence was found of non-neighbour novel words enabling re-tuning of phoneme boundaries through perceptual learning when the novel words were trained without meaning. Figure 9 shows that there was a small phoneme boundary shift for participants who learned /t/-ending novel words (non-significant on day 1, significant on day 2), and a somewhat larger and significant shift in participants who learned /d/-ending words. However, this latter shift went in the wrong direction:

these participants made more /t/-responses on day 1 and day 2 compared to baseline, even though they were exposed to the ambiguous phoneme in a novel lexical context supporting a /d/-interpretation. This suggests that the tendency to shift towards responding /t/ in the non-neighbour group was insensitive to the lexical manipulation using novel words, and thus does not constitute evidence that these novel words show lexical integration. Figure 10 shows the net effect of the two bias groups: there was a 3% change from baseline on day 1, and a 4% change after a further training session on day 2, both in the wrong direction.

The situation was very different looking at neighbour novel words. Figure 8 shows a shift in both continua going in the direction predicted by the lexical bias afforded by the novel words. Figure 10 shows there was an 8% change from baseline on day 1, and a 7% change on day 2. The change on day 1 was statistically significant, while the change on day 2 approached significance ($p = .07$). The magnitude of these changes is quite comparable to the data reported by Leach and Samuel. They reported two experiments using the same design as the experiment reported here, i.e. a contrast between pre-exposure phoneme categorisation and post-exposure categorisation, both carried out on the same day. In a semantic training condition they reported a shift of 13% on a /s/ - /sh/ continuum with /s/-ending novel words, and a 7% shift with /sh/-ending novel words. In a non-semantic experiment they reported non-significant shifts of 3% and 4%, in the wrong direction. The main experiments in Leach and Samuel did not use a baseline contrast, so the magnitude of the effect in those experiments cannot be directly compared to the effects in the current experiment. It seems then that the effect observed here was very similar in magnitude to the effect seen by Leach and Samuel. The small advantage in their favour might be explained by the different continua used in the two experiments, the /s/ - /sh/ continuum may lend itself better to an ambiguity manipulation that the /t/ - /d/ continuum.

The current data replicated the Leach and Samuel finding with respect to their time course too. When perceptual learning was observed, it emerged immediately after training. In fact, the effect appeared to be weaker when tested a day later, in spite of the additional training provided on day 2. These data together with data from Experiment 2 support the idea that perceptual learning does not seem to benefit from offline consolidation.

The striking difference between lexical integration in neighbours and non-neighbours is not present in the measure of lexical configuration assessed here by cued recall. Although there was a small numerical advantage for neighbours on both days, the difference failed to reach significance, suggesting that the lack of perceptual learning effect in non-neighbours was not due to these items being more difficult to learn. Instead, the more likely explanation is that since neighbours overlap with existing words, they can evoke the meaning of their closest neighbour, and benefit from this semantic support to a degree significant enough to allow for lexical integration to take place. The cued recall task also showed increasing accuracy on day 2, however this is not surprising as day 2 also included another training session providing further exposure to the novel words.

2.5 Chapter Summary and General Discussion

The motivation for the series of experiments described in this chapter was the finding reported by Leach and Samuel (2007) that only meaningful novel words showed evidence of lexical integration when measured by perceptual learning. These authors found no lexical integration when the novel words were trained in a phoneme monitoring task. However, when the same stimuli were trained in tasks which provided a meaning for the words, lexical integration did emerge. Measures of lexical configuration, that is, explicit knowledge of the form of the words, were unaffected by the semantic manipulation. This study provided strong evidence that meaning is necessary for novel words to develop lexical representations that are fully functional and engage in word-like behaviours at sublexical and lexical levels.

This was a surprising finding in light of the lexical competition data provided by Gaskell and colleagues, which have consistently shown novel words engaging in lexical competition with existing words as a result of non-semantic training. The one study which directly contrasted semantic and non-semantic training (Dumay et al., 2004) showed no benefit of semantics on novel word lexical competition.

The hypothesis presented in this chapter was that the reason why meaning has not been implicated as an important factor in the lexical competition studies is that the stimuli have been neighbour novel words. These novel words overlap to a large extent with existing words in the lexicon. Hence they may also evoke the meaning of the base words from which they were derived, and this “inherited”

meaning may be enough to allow for lexical integration to occur. Experiment 1 provided support for this notion. Neighbour meanings were easier to learn than non-neighbour meanings, but only when the neighbour meaning was related to the meaning of the base word. This suggests that the base word meaning did exert an influence on the learning process, facilitative in the case of consistent neighbours, and less useful in the case of inconsistent meanings. Participants also recalled more neighbour word forms than non-neighbour forms. This could be explained simply by pointing out that neighbour forms had the extra mnemonic support from their base word forms. However, the semantic consistency manipulation played a role even in this non-semantic task: more consistent neighbour word forms were recalled than inconsistent forms.

Once Experiment 1 had established that the base word meanings are available to participants in a word learning experiment, and that they appear to influence the degree to which novel word meanings and forms are encoded, Experiment 3 sought to demonstrate that base word meanings can be influential in the emergence of lexical integration as well. In an experiment modelled after the Leach and Samuel (2007) studies, I showed that neighbours do in fact show lexical integration even when they are trained using a non-semantic task. Non-neighbours showed no such effect, replicating the failure of Leach and Samuel to see lexical integration with meaningless novel words that do not overlap with existing words. Taken together, the series of experiments by Leach and Samuel and the present experiments suggest that meaning is important in novel word learning, and that novel words that overlap with existing words, in the absence of experimentally trained meaning, can evoke the meaning of the base words and solve the apparent discrepancy between the perceptual learning data and the lexical competition data. This is further supported by the priming studies reviewed in the introduction to this chapter, which showed that nonwords derived from real words can prime real words that are semantically related to the base words.

There is another variable that may have contributed to the learning of the neighbours. As mentioned in Chapter 1, Storkel et al. (2006) have suggested that phonetic neighbourhood density affects word learning in children and in adults. According to this view, adding a novel word to a high-density neighbourhood is easier because hearing the novel word will activate many neighbours, with the activation feeding back to the phonological level, and back again to the lexical level

resulting in more co-activation of phonemic and lexical representations than would be the case for novel words in sparse neighbourhoods. The neighbour novel words used in the current experiments would have by virtue of their design fallen into denser neighbourhoods than the non-neighbours. This may have contributed to the emergence of lexical integration in neighbours. Leach and Samuel argued along the same lines when they suggested that one advantage that neighbour words have is that they may not require building a new lexical representation from scratch.

The neighbourhood and semantic accounts are unlikely to be mutually exclusive. Both are probably important in word learning (see e.g. Storkel, 2009, for a demonstration of both phonological and semantic predictors of infants' word knowledge), and they may interact. For example, it is possible that meaning is necessary particularly when adding lexical entries in sparse neighbourhoods, as was the case in Leach and Samuel (2007). I am not aware of studies that would have manipulated both variables in the same experiment. Storkel et al. (2006) manipulated neighbourhood density but used meaningful training (story context) only. Leach and Samuel (2007) and Dumay et al. (2004) varied training but used the same stimuli throughout their experiments. It is left for future studies to tease apart the effects of the two explanatory variables in adult word learning.

Experiment 3 was successful in both replicating the Leach and Samuel (2007) finding of no lexical integration with non-neighbours in training with no meaning, and in extending these findings by showing that lexical integration is found with neighbours under the same circumstances. The second way in which their original data were replicated was in terms of the time course of the effect: they found the perceptual learning effect immediately after training, as was the case in the current experiment too. Leach and Samuel did not discuss why the perceptual learning measure allows lexical integration to be seen immediately, while lexical competition effects tend to benefit from a delay between training and test. As discussed in Chapter 1, Davis and Gaskell (2009) suggested this may be due to the properties of complementary learning systems and early lexical representations having direct access to a phonological level of representation.

Finally, while the main purpose of Experiment 2 was to generate a set of materials suitable for perceptual learning, it also revealed that the phoneme categorisation boundary shift induced by perceptual learning with real words is robust against decay over a week. This is an impressive demonstration of flexibility

in the speech recognition system, showing that even though participants must have been exposed to countless tokens of /t/ and /d/ produced by a variety of speakers between the day 1 and day 8 sessions, they still retained the adjustment made in response to the ambiguous /?dt/ heard on day 1. Whether this would also be the case for boundary shifts created by novel word lexical contexts remains to be shown.

It is issues of time course of novel word learning that I will turn to in the following chapters. The experiments reported in this chapter strongly support the view that meaning is important in novel word learning, both in looking at explicit measures of lexical configuration, and implicit measures of lexical integration. While we have access to data looking at offline consolidation effects in word form acquisition (e.g., Dumay & Gaskell, 2007), there are very little data on consolidation of semantic information in novel word learning. Experiment 1 provided preliminary data on this issue. In that experiment it seemed that explicit recall of novel word meanings decays over the first 24 hours, while explicit knowledge of word forms remains unchanged. This dissociation between meaning and form recall will be further discussed in later chapters (recall also the finding by Clay et al. suggesting that automatic semantic access benefits from consolidation). Chapter 4 will introduce novel ways of examining consolidation of form and meaning knowledge. Chapter 5 will focus on the role of offline consolidation in different types of access to novel word meaning. Chapter 6 will return to the issue of form knowledge, and attempt to find out which aspects of sleep architecture drive the consolidation effect in novel word learning. Prior to moving on to the experimental chapters, the next chapter will review literature on offline consolidation and the role of sleep in this process, with emphasis on learning linguistic materials.

Chapter 3: Memory consolidation, sleep, and language learning

3.1 Origins of consolidation theory

The first ideas about memory consolidation can be traced back to two significant early findings. Ribot (1882) described a number of patients suffering from retrograde amnesia. These patients had typically lost memory of events that occurred shortly prior to the brain insult that triggered the onset of amnesia, but could still recall many events from the more distant past. This led Ribot to suggest a time-dependent process of memory organisation, whereby newly acquired memories become "fixed" over time and gain resistance to trauma (Polster, Nadel, & Schacter, 1991). About twenty years later Mueller and Pilzecker (1900) provided experimental evidence of a consolidation process in healthy adults (see Lechner, Squire, and Byrne, 1999, for a summary of this work). In these experiments participants were asked to learn a list of nonword pairs within a fixed time, with memory tested after a brief delay by asking participants to produce the appropriate nonword when cued with the other member of a pair. Mueller and Pilzecker noted that if learning of the initial list was followed by learning of another list, recall on the first list was significantly impaired compared to a condition where a training-test interval of identical length did not include intervening learning. However, if a gap of about 6 minutes was allowed between presenting the two study lists, no impairment was seen in a later recall test of the first list. Mueller and Pilzecker concluded that reading the initial study list engaged memory-related processes which continued to strengthen the memory trace for several minutes after the study session and could be disrupted by introducing interfering material before consolidation was complete.

Similar ideas were subsequently tested in a wide research effort in the first half of the 20th century. The phenomenon which served as the focus of these studies was termed "reminiscence", defined as an improvement in recall over time of materials learned but not rehearsed again prior to testing. Buxton (1943) in his classic review of this work highlighted among other concerns the difficulties of replicating many of the reports of reminiscence, the problems with confounds in experimental design (many studies had failed to control for the potential additional learning occurring as a result of repeated testing), and the restricted nature of the

stimuli typically used at the time (most often nonword syllables or poetry). It may be that these problems were partially responsible for the declining interest in this line of research within the cognitive tradition during the following decades, with few modern cognitive formal theories of memory including a role for consolidation (Brown & Lewandowsky, in press).

In contrast to memory research in cognitive psychology, the idea of consolidation is broadly accepted in cognitive neuroscience (although see Moscovitch, Nadel, Winocur, Gilboa, and Rosenbaum, 2006, for a sceptical view). Hebb (1949) was the first to present a biologically plausible mechanism for consolidation in short term memory. He acknowledged the contradiction between the need to lay down new memories quickly as a result of limited exposure to the stimulus to be learned, and the slow rate of structural changes which are needed for permanent memories to form in the brain. His solution was to propose a dual trace mechanism, which relied on the notion of reverberatory action within cell assemblies. Reverberation allowed activation of cells to persist after the offset of the stimulus at least long enough to allow the connections between the neurons to strengthen. Reverberation however only spans timescales of seconds or less, and is an example of short-term consolidation, now often referred to as cellular consolidation, thought to involve biochemical reorganisation of relevant synapses. Marr (1970, 1971) proposed a neural theory of memory formation which included a consolidation process that occurred over days. According to this theory, memories are initially stored in “simple memory”, whose neural basis is the archicortex (part of the cerebral cortex that includes the limbic system). Simple memory however does not allow information to be classified or generalised with relation to existing memories. Information needs to be transferred from simple memory to the neocortex for these processes to take place. Marr suggested that this transfer occurs mostly during sleep. This type of long-term consolidation is sometimes referred to as systems-level consolidation as it involves more significant structural change. Most of Marr’s basic ideas were very similar to modern views of memory consolidation and the complementary learning system approaches I briefly described in Chapter 1, including the role for sleep they postulate. I will next discuss the modern theories and the supporting evidence for these theories, much of which comes from sleep research.

3.2 Modern theories of complementary learning systems

The two currently most influential theories of complementary learning systems (CLS) were put forward by Squire (1992) and by McClelland et al. (1995), with the former elaborated by Alvarez and Squire (1994), and the latter later further developed by O'Reilly and Norman (2002). These theories were largely motivated by demonstrations both in humans and animals of the importance of the hippocampal formation and related medial temporal areas in amnesia. For example, Zola-Morgan and Alvarez (1990) found that monkeys showed a typical Ribot gradient of retrograde amnesia after surgical removal of the hippocampal formation. This was a similar deficit to what was seen in patient HM who had undergone surgery during which large parts of his hippocampal system was removed (Scoville & Milner, 1957). These observations strongly suggested that the hippocampus plays a temporally limited role in the encoding of new memories. One advantage that was not available to Marr but which is used by the modern CLS theories is computational modelling. The modern theories mentioned above have been implemented as connectionist models, increasing the detail and predictive power of the CLS framework (Alvarez & Squire, 1994; McClelland et al, 1995). Although the models differ in their detailed implementation, their basic principles are essentially the same, so the brief description given in Chapter 1 and the more detailed description below apply to them all, unless otherwise stated (see also Meeter and Murre, 2005, for a recent model).

CLS models postulate two learning systems with different rates of learning, different structures of representation, and a transfer process between the two systems. The neocortex is responsible for permanent memory storage, and information here is represented in a form of overlapping representations. This overlap allows the information to be integrated with existing knowledge and generalisation in learning to take place. However, there are several reasons why the learning rate needs to be slow in a system like this. McClelland et al. (1995) argued that slow, interleaved learning is necessary to ensure that the existing knowledge structure is not disturbed by the incoming information. This is a particularly important lesson from computational modelling, where massed learning of new information has been shown to lead to catastrophic interference where the new information can overwrite the old information (although see Page, 2000, and the

following commentaries to his target article for an argument that catastrophic interference is only a problem for certain types of computational models, i.e. distributed models). Alvarez and Squire (1994) further motivate a slow learning rate by pointing out that memories in the neocortex are represented over geographically disparate areas with slowly emerging connectivity. In these models consolidation occurs in the form of gradual strengthening of the neocortical trace, during a reinstatement process where the memory trace is repeatedly re-activated over time. Reinstatement can occur with direct exposure to the stimulus, or with explicit recall of the stimulus. However, as recognised by the early memory theorists, much of the time new memories are generated in response to events that occur only once, not providing the necessary reinstatement. Hence the CLS models argue, following Marr (1970, 1971), that reinstatement takes place offline as well, including during sleep.

Because the initial neocortical memory trace is too weak to activate fully on its own in the absence of external stimulation or in response to partial stimulation, a second system is needed to "bind" the neocortical representation together until it has become strong enough through consolidation. The CLS models suggest that this system is provided by the medial temporal lobe, including the hippocampus. Unlike the neocortex, the hippocampus allows fast generation of sparse, non-overlapping representations. The sparsity of the hippocampal representations circumvents the catastrophic interference problem while allowing fast learning. Furthermore, since the connections between the hippocampus and neocortex can be modified quickly, the hippocampal and the emerging neocortical representations are linked immediately. This means that even partial activation of the emerging neocortical representation can be boosted with the support from the hippocampus to activate the complete representation. As the neocortical trace gains in strength with consolidation, the hippocampal support will eventually become superfluous, and the neocortical memory will eventually become independent of the hippocampus.

The CLS models are supported by a rich literature from neuroscience. As outlined earlier, they can explain the Ribot gradient in amnesia by referring to the role of the hippocampus in unconsolidated memories which become independent of the hippocampal formation with time. There is also an increasing literature from animal and human studies that demonstrates hippocampal involvement in accessing newly acquired memories, and reorganisation of these memories over time. A thorough review of this literature is beyond the scope of this thesis, but two

complementing animal studies give a good flavour of the data. Wirth et al. (2003) trained monkeys to associate photographic scenes with specific spatial location within the scene. Activity of a set of hippocampal cells was recorded during the training. The recordings showed a stimulus-selective firing rate change whose emergence closely coincided with behavioural learning, suggesting a rapid change in hippocampal response properties. In contrast, Takehara-Nishiuchi and McNaughton (2008) recorded activity from rats' medial prefrontal cortex cells during a conditional associate learning task. In this neocortical area a firing pattern sensitive to the conditioning manipulation only emerged after several days of training, and further increased offline over a period of several weeks, demonstrating a consolidation process in action. In human studies the data are more equivocal, as some neuroimaging studies have found hippocampal activation during retrieval of both remote and recent memories, while others have found it only with recent memories (see Meeter and Murre, 2004, for a review). Recent advances in imaging brain connectivity are enabling more sophisticated data to be gained though. Takashima et al. (2009) trained participants to associate faces with spatial locations, and tested recall immediately or 24 hours later while monitoring hippocampal and neocortical activity using fMRI during the test. Comparison of the immediate and delayed testing showed that hippocampal activity decreased while neocortical activity increased during consolidation. In addition, connectivity between the hippocampus and neocortex decreased over time, while connectivity within neocortical representational areas increased. Taken together these studies offer powerful evidence in favour of the consolidation proposed by CLS models.

3.3 Memory consolidation and sleep in language learning

Both Marr (1970, 1971) and McClelland et al. (1995) suggested that much of the reinstatement of newly acquired hippocampally-mediated memories occurs during sleep. If this was true, then it should be possible to see better memory performance after retention periods involving sleep in contrast to equivalent periods of wakefulness. Such data were indeed reported at an early stage. Van Ormer (1933) in an early review of the literature looking at sleep and memory cited work carried out by Ebbinghaus in 1885, where he saw better retention of newly learned materials over a period of time involving sleep than what would be predicted by mere passage

of time. This literature has vastly expanded in the past decade, with most evidence of performance improvements over sleep coming from procedural tasks, typically studied by looking at motor skills (e.g., Walker & Stickgold, 2006). The recent years have also seen several reports of sleep benefits in tasks measuring declarative learning, often looked at by measuring word-pair learning (e.g., Marshall & Born, 2007). While such demonstrations of sleep benefits are compatible with CLS theories, direct evidence for reinstatement during sleep is also now available. For example, Peigneux et al. (2004) showed that areas of the human hippocampus that were active during a route learning task were reactivated during subsequent sleep (but see also Tononi and Cirelli, 2006, for a view of sleep and memory that does not include reinstatement).

The idea that sleep plays a crucial role in memory consolidation is not universally accepted though. Siegel and Vertes for example have made a number of points that seem to undermine the hypothesis (Vertes & Eastman, 2000; Siegel, 2001; Vertes, 2004; Vertes & Siegel, 2005). One of these claims is that sleep does not consolidate declarative memories. Although it is true that there are more reports of sleep effects in the procedural domain, there is now a large number of studies looking at declarative memory too, although most of these use only word-pair learning (Plihal & Born 1997, 1999; Gais & Born, 2004; Marshall, Helgadottir, Molle & Born, 2006; Marshall & Born, 2007). A related criticism states that even in the procedural domain, the literature is inconsistent in that some studies have failed to find a sleep effect, and those that have found it do not consistently implicate the same sleep stages driving the effect (see Vertes, 2004, for a detailed discussion). A third point concerns individuals in whom REM sleep is suppressed or eliminated due to use of antidepressant drugs or brainstem lesions. These people are able to live normal lives without any apparent learning difficulties. However, as pointed out by Walker and Stickgold (2004), these populations have not actually been tested on the tasks shown to be affected by REM sleep. Also, there is now ample evidence to show that REM is not the only stage that is involved in memory formation; slow-wave sleep and stage 2 sleep have also shown to play a role (Diekelmann & Born, 2010). Finally, Siegel (2001) has argued that if REM sleep were indeed necessary for memory, then those mammals with the highest intelligence should show the highest amounts of REM sleep. This turns out not to be the case, as humans fall into the average range in amount of REM (Siegel, 2001). This argument however assumes

that sleep currently serves the same evolutionary purpose in all mammals, which is unlikely to be true (e.g., Horne, 2006).

Although the criticisms mentioned above cannot be dismissed, there is now an abundance of evidence supporting the view that sleep both stabilises and enhances new memories (Walker, 2005) from a wide range of memory domains in addition to procedural and declarative learning, including studies into sleep's protective effect against interference from competing novel memories (Ellenbogen, Hulbert, Stickgold, Dinges, & Thompson-Schill, 2006), relational memory (Ellenbogen, Hu, Payne, Titone, & Walker, 2007), emergence of insight into implicit rules in complex tasks (Wagner, Gais, Haider, Verleger, & Born, 2004), creativity (Cai, Mednick, Harrison, Kanady, & Mednick, 2009), and emotional memory (Walker, 2009). A detailed description of this literature is not attempted here, instead I will now turn to studies that have specifically looked at consolidation and sleep in language learning. These studies cover several levels of language processing, from phonological to lexical, syntactic, and semantic levels of processing. I will follow a bottom up approach in outlining this body of research below.

3.3.1 Consolidation of phonological, lexical, and syntactic knowledge

Fenn et al. (2003) trained participants to recognise computer-generated speech. Half of the participants (wake group) were pre-tested and trained in the morning, and retested on a new set of stimuli produced by the same speech generator after 12 hours in the evening. The other group was trained in the evening, and had their post-test after 12 hours in the morning (sleep group). Note that the wake group had no sleep between training and the post-test, while the sleep group did. Compared to pre-test, the sleep group showed a larger improvement in performance than the wake group (18% vs. 10%). Control groups that were tested immediately after training showed an improvement similar to that of the sleep group, suggesting that staying awake between training and test conferred a cost on retention of the phonetic learning gained during training. Fenn et al. concluded that sleep consolidates new phonetic perceptual skills in two ways: firstly, it protects the new memories against interference (or decay) during the day. Secondly, sleep appeared to recover skills lost during the day. The latter claim was motivated by performance in one of their control wake groups which showed the decline in performance after a day of

wakefulness, but improved significantly in a second post-test taking place after a night of sleep.

Sleep appears to be relevant in learning novel words too. One study has shown this to be the case in learning words from a foreign language. Gais, Lucas, and Born (2006) asked English speaking participants to memorise English-German word pairs. Recall of the German words in response to English prompts was tested immediately and again 24 or 36 hours later. Half of the participants were trained in the evening, so they got to sleep immediately after learning. The other half were trained in the morning, so training was followed by a normal day of wakefulness prior to a normal night of sleep. Participants who slept shortly after training forgot significantly fewer words than participants who were trained in the morning. A further experiment controlled for circadian effects by training all participants in the evening. Half of the participants proceeded to have a normal night of sleep, while the other half were sleep deprived for the duration of the night. Recall test after a recovery night showed that the sleep group forgot significantly fewer words. The authors argued that consolidation of novel vocabulary benefits from sleep most when the time interval between learning and onset of sleep is short. In this study the participants had no previous knowledge of the German language, so it may be viewed as examining the very early stages of learning new language vocabulary. De Koninck, Lorrain, Christ, Proulx, and Coulombe (1989) looked at slightly more advanced L2 learning by taking polysomnographic (PSG) measures of students participating is a six-week French language immersion course prior to, during, and after the course. These native English speaking students showed a positive correlation between learning progress and increase in rapid eye movement (REM) sleep during the course, suggesting that at least this stage of sleep is associated with language learning. The possible roles different sleep stages may have in language learning will be discussed further in Chapter 6, in the current chapter I limit the discussion to the global effects of sleep and consolidation on learning. Note also that a language immersion course is not restricted to word learning, so it is impossible to say which aspect of language acquisition this study is most relevant to.

Acquiring words in the learner's own language has recently attracted interest, and much of this literature was already reviewed in Chapter 1. The data most relevant for consolidation come from the lexical competition studies reported by Gaskell and colleagues (e.g., Gaskell & Dumay, 2003). Recall that these studies

have shown that novel spoken words engage in lexical competition with overlapping existing words only in a delayed competition task, suggesting a role for consolidation. Bowers et al. (2005) showed this to be the case for written words as well. Dumay and Gaskell (2007) further showed that the auditory lexical competition effect can emerge already after 12 hours, provided that sleep has occurred during that time. Such sleep-dependent consolidation fits easily within the CLS framework. Davis et al. (2009) provided evidence in favour of this interpretation. The main neocortical findings of their fMRI experiment were described in Chapter 1, but it is worth drawing attention here to their hippocampal data. When a contrast was made between consolidated, unconsolidated, and untrained novel words, a region of interest encompassing the hippocampus showed higher activation to untrained than unconsolidated words in the first scanning run. This suggested that the untrained novel words, which had never been heard before, engaged the hippocampus upon their first exposure. This activity declined in the following two scanning runs, and became similar to activity elicited by trained novel words. Magnitude of hippocampal activity on the first run was also positively correlated with later recognition memory of the novel words in a 2AFC task.

A similar pattern of hippocampal activity was reported by Breitenstein et al. (2005). In this experiment participants learned novel names for familiar objects while brain activity was monitored using fMRI. As training progressed, hippocampal response to the novel words decreased. This hippocampal disengagement was mirrored by increasing activity in the inferior parietal lobe as a function of training. The authors suggested that this is the neocortical site for permanent phonological storage of novel words. Hippocampal activity during training also correlated with behavioural measures of novel word knowledge. Participants who showed less decrease of hippocampal activity over the course of training had better learning outcome at test, and improved more during training. Hippocampal activity at the first training block also predicted learning outcome, perhaps indicating that hippocampus-dependent episodic memory of the novel words was helpful at test. Unfortunately this study did not include a re-test after sleep, but does suggest that a consolidation process involving hippocampal-neocortical interaction is initiated from the very beginning of word learning. It could be argued that the reduction in hippocampal activity as training progressed was due to adaptation. However, a control condition where novel words and object pictures were randomly paired from trial to trial

showed no such effect. Adaptation also cannot explain the dissociation between hippocampal disengagement and increasing parietal activity.

Gomez, Bootzin, and Nadel (2006) sought to establish whether sleep helps infants to learn syntactic structure in an artificial language. 15-month-old infants were exposed to sentences from a language which had syntactic structure based on nonadjacent word pairs. The sentences consisted of three words, such as *pel-wadim-jic*. Each infant heard two types of sentences where the first word always predicted the third word while the middle word was unpredictable (e.g., *pel-X-jic* and *vot-X-rud*). After a nap (sleep group) or an equivalent time awake (wake group) infants were tested by exposing them to the trained sentences and new sentences with the same predictive rule while monitoring their gaze direction. The wake group looked more towards the direction where they heard the trained nonadjacent word pairings compared to unfamiliar pairings, indicating that they had an accurate memory of the trained items. In the sleep group however looking preference was not determined by the training items, but rather by the first sentence the infants heard in the test session. In the following trials infants preferred sentences that conformed to the first sentence. It seemed that while the sleep group showed no evidence of recalling the specific training items, they were able to apply the rule established by the training items to guide their processing of novel test items, a demonstration of grammatical rule abstraction the wake group failed to show. The effect was also shown to be robust after a 24 hour delay, but could only be observed if sleep occurred within 4 hours of the initial exposure (Hupbach, Gomez, Bootzin, & Nadel, 2009). A similar finding was reported by St. Clair and Monaghan (2008), who showed that adults learning an artificial language could categorise words in the language into grammatical categories based on phonological and distributional cues inherent in the language. Importantly, only participants who slept between training and test showed generalisation of this ability to words that had not been included in the training. It is striking that in both of these studies the sleep group showed no evidence of recalling the specific training materials, but were able to apply the rules inherent in the training to new stimuli. The wake groups on the other hand showed the opposite pattern. This suggests that sleep is particularly useful in generalisation and abstraction of newly learned linguistic information.

3.3.2 Consolidation of semantic knowledge

The studies reviewed above suggest that memory consolidation is an important factor in sublexical, lexical, and syntactic learning, and that the consolidation process benefits from sleep in all of these cases. The final level of language learning I will discuss in this chapter is learning the meaning of novel words. Much of the relevant literature was already reviewed in Chapter 1. Recall that both Perfetti et al. (2005) and Mesters-Misse et al. (2007) used a primed semantic decision task where a newly learned word acted as the prime in an ERP experiment. In the Perfetti et al. study priming was seen in reaction times (faster responses in trials involving a semantically related novel word compared to unrelated trials), and in the amplitude of the N400 (higher amplitude in the unrelated condition) immediately after training. This might be interpreted as showing immediate semantic learning in the absence of consolidation, however there are two points that urge caution with regard to this study. Firstly, the target words in the related condition included words that had occurred as part of the definition of the novel word (the exact proportion was not given: "many trained words were paired with a meaning probe that had occurred as part of the definition", Perfetti et al., 2005, p. 1282). This means that in many of the trials the priming effect was potentially episodic rather than purely semantic. Secondly, since the semantic decision task requires an explicit judgement to be made about the meaning of the novel word, it is not clear whether that task measures semantic priming in the same sense as the more commonly used priming tasks only requiring explicit processing of the target. Mestres-Misse et al. (2007) also used a semantic decision task, but did not observe a behavioural priming effect. In fact, here the opposite pattern was seen, possibly reflecting poor explicit learning of the novel word meanings. They did however observe the N400 effect with the novel words. Interestingly though, the N400 was not identical to a real word control condition: with the novel words the N400 latency was delayed, and it had a different source. A typical N400 parietal source was seen for real words, but the novel word condition revealed a frontal source, possibly reflecting a more effortful semantic retrieval strategy. Mestres-Misse, Camara, Rodriguez-Fornells, Rotte, and Munte (2008) replicated this ERP study using fMRI to identify the brain regions involved in semantic word learning. This time a behavioural priming effect in the semantic decision task was obtained, although in this version of the task the novel

word prime was followed by the actual meaning of the novel word, making it again difficult to say whether semantic or episodic priming was being measured. The imaging data showed activation of several neocortical areas in response to the novel words, but interestingly also more activation of the medial temporal lobe areas in response to novel words than real words, again suggesting hippocampal involvement in the early stages of word learning.

Breitenstein et al. (2007) argued that newly learned meaningful words show cross-modal priming effects (see Chapter 1). This was taken as evidence that the new words were integrated in existing language networks. However, the priming test was carried out after five days of training, excluding the opportunity to evaluate a potential role for consolidation. Furthermore, these priming data were confounded with a possible response congruity priming effect: the target required a living/non-living decision, and the prime was always response-congruent with the target. Also, the novel word-picture pairs used at test were the same pairs as used in training, again making it difficult to distinguish between semantic and episodic priming accounts. The MEG follow up to this behavioural work (Dobel et al., in press) alleviates these concerns to some degree by showing an N400m attenuation to novel word priming trials from pre-training test to post-training test, if the N400m is taken to be a component sensitive to semantic processing (only trials with identical or semantically related prime-target pairings were used here). It should also be noted that in the post-training test, the N400m to novel word trials was still significantly larger than in a control condition using an identical prime real word – target picture pairing (although not significantly different from a control condition using semantically related real prime word – target picture pairings). Like the subtle differences in the N400 latency and source to novel word trials compared to real word trials reported by Mestres-Misse et al. (2007), this might reflect an incomplete consolidation process.

An earlier priming study using newly learned words was reported by Dagenbach, Horst, and Carr (1990). Here participants were taught the meanings of a list of rare words (e.g., *drupe*, which is a botanical term for a fleshy fruit). After the meanings of the new words had been reliably learned, participants were asked to memorise an episodic study list of word pairs combining a novel word and a synonym or a semantically related familiar word (e.g., *drupe – cherry*). These two study phases were followed by a test of priming, where lexical decisions were made

to the familiar words, which could be primed or unprimed by the novel word they were paired with in the episodic training. No episodic priming was found immediately after the training session. The authors suggest this was due to insufficient training, and carried out another experiment where participants trained on the novel word meanings and episodic pairs for five weeks. At the end of the five weeks priming was found. It is not possible to say whether this delayed effect emerged due to the additional training, or to the delay between beginning of training and the test session, as might be predicted by consolidation theory. However, it does seem that this priming effect was not immediate. A further complication is caused by the relationship between the prime and the target, which were both semantically and episodically related.

Finally, the picture-word interference work by Clay et al. (2007) is highly relevant here, as discussed in Chapter 1. Recall that in this study a non-semantic PWI effect was found immediately after training whereby a superimposed novel word interfered with the naming of a picture, whether or not the word and the picture were semantically related. However, in the second testing session which took place one week later, the semantic PWI effect was also seen, with a semantically related novel word interfering with picture naming more than an unrelated novel word. This is strong behavioural evidence that automatic semantic activation of novel words is not seen until after some period of consolidation has taken place. The exact necessary duration of consolidation remains to be defined.

3.4 Conclusions

I have reviewed evidence in this chapter for the role of memory consolidation and sleep in various aspects of language learning. The current evidence appears to be strongest in the case of learning novel word forms, in particular when the integration of those words in the mental lexicon is considered. Most crucially for the work presented in the next two chapters, the current state of research with regard to consolidation of novel word meanings is less clear. Most of the work discussed in the previous section has only looked at learning immediately after training. Although some of those data gave the first impression that semantics does not require consolidation (e.g., Perfetti et al.'s priming data), a closer look revealed both differences in how familiar and novel word meanings were processed at this early

time point (e.g., Mestres-Misse et al.'s ERP data) and concerns about tasks and methodology.

There is however good reason to believe that offline consolidation and/or sleep play a role in learning novel word meanings. Walker (2009) has argued that sleep may be crucial in integrating and relating new memories with existing memories. This is not only supported by the lexical competition data (Dumay & Gaskell, 2007), but also by data from a range of different tasks. Ellenbogen et al. (2007) had participants learn premise pairs (e.g., A > B, B > C, C > D, D > E, E > F) and tested their recall of the pairs immediately and after a delay. While all participants had good recall of the pairs at both testing times, only in the delayed test was there any evidence of participants having interrelated the new information into a hierarchy (A > B > C, etc). Furthermore, when looking at long-distance item separation (e.g., interrelating B with D, to give B > D), participants who had slept between training and the delayed test showed significantly better learning than participants who had remained awake. Relevant data were also presented by Payne et al. (2009) in an experiment looking at false memory creation. These authors used the Deese-Roediger-McDermott (DRM) paradigm where participants learn lists of semantically associated words, and typically during recall present with false memories by adding new untrained related words to the lists. Payne et al. showed that participants who slept between training and recall showed more false memories than participants who remained awake, suggesting that sleep helped to relate the study words with associated unstudied words [5]. These two studies suggest that offline consolidation not only enhances memory of the trained materials, but also helps to integrate them with other trained stimuli as well as existing knowledge. This implies that learning novel meaningful words may also benefit from consolidation, especially when measured in tasks that require the novel meanings to be related with existing word knowledge, such as semantic priming.

In order to try to clarify this state of affairs, the next two chapters will look at different types of semantic priming and recall in novel words, both immediately after training and after a delay of at least one day (semantic decision in Chapter 4,

[5] Fenn, Gallo, Margoliash, Roediger, & Nusbaum (2009) found the opposite effect, with sleep reducing false recall. The pattern of data reported by Payne et al. (2009) however has recently been replicated by a different research group, suggesting it is reliable (J. D. Payne, personal communication, December 11, 2009)

semantically primed lexical decision in Chapter 5). The time course of semantic consolidation will also be contrasted with that of word form consolidation using novel tasks (Chapters 4 and 6).

Chapter 4: Consolidation in learning novel word meanings and forms

4.1 Introduction

Experiment 3 showed that meaning is indeed crucial for at least some aspects of lexical integration. Only those novel words which potentially inherited the meaning of their base words showed the ability to reconfigure phoneme boundaries. As meaning appears to play such a major role in novel words integrating in the lexicon, it is important to understand better the process through which novel words link with semantic representations, and whether this process benefits from offline consolidation in the same way as novel word forms seem to. Clay et al. (2007) showed in a picture-word interference task that novel word meanings do appear to benefit from consolidation, at least over the course of one week. However, as that study did not include re-test sessions earlier than the one-week follow up, it is unclear whether semantic consolidation operates on the same time scale as word form consolidation (where the first sleep period appears to be crucial), or whether semantic consolidation possibly is a more temporally drawn out process, operating over several days and/or nights. In this chapter I shall report two experiments, the first of which focused on acquisition of novel word meanings, and the second focusing on acquisition of novel word forms, evaluating consolidation within the first 24 hours. The main task looking at consolidation of meaning here is semantic decision. As reviewed in the last chapter, this task has been used before by Perfetti et al. (2005), and Mestres-Misse et al. (2007, 2008), but these studies involved some methodological issues that complicated their interpretation (see Chapter 3). Experiment 4 attempted to circumvent these methodological issues, and also added a new task looking at semantic access: sentence plausibility judgement.

4.2 Experiment 4

To evaluate the time course of novel word learning, Experiment 4 used the training and testing schedule introduced by Davis et al. (2009). Participants learned one set of novel words on day 1 (consolidated), and another set on day 2 (unconsolidated), and were tested immediately after the day 2 training session. Hence only the first set of novel words had a chance to benefit from offline

consolidation during the 24 hours between training and testing. In the test session participants were required to make semantic decision responses to prime-target pairs where the prime was a novel word. In the sentence plausibility judgement task a decision was made about the appropriateness of using a given novel word in a sentential context. Finally, explicit knowledge of novel word meanings was assessed by meaning recall. If novel word meanings benefit from offline consolidation, we should see a larger priming effect in semantic decision and faster RTs to consolidated than unconsolidated words. The same pattern should be seen in the sentence task. Data from Experiment 1 however suggest that explicit recall of the meanings is likely to be unaffected by consolidation. A single-word shadowing task was also included to evaluate form-based consolidation in the same experiment.

4.2.1 Method

Materials

The novel word stimuli consisted of 102 written pronounceable nonwords (e.g., *feckton*) created by Deacon et al. (2004), with average length of 6.5 letters (SD = 0.9). These nonwords were not derived from real words and were designed not evoke the meanings of real words, as confirmed by a norming study by Deacon et al. (2004). Experiment 3 showed that novel words that are closely related to real existing words may activate the meaning of the real word. Hence it was important here to choose novel word forms where there was little chance of interference from existing word meanings. These nonwords are presented in Appendix 4.

In this experiment each novel word was coupled with a meaning during training. A novel word meaning was represented by a familiar noun referring to an existing object (e.g., "feckton is a type of *cat*"). The novel word meanings were selected from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 2004). This is a corpus of 5019 words for which a large number of participants have provided free association responses. Each word in the corpus has a list of associated words, and each associate has a numerical association strength value based on the proportion of people who produced that associate as a response to the stimulus word. Sixty-eight words were chosen from this set to act as the meanings assigned to the novel words. All selected meanings were nouns of medium frequency (M = 73.7) with three strong associates (where a noun had more

than three strong associates in the corpus, the three strongest ones were selected while avoiding overlap with other associates). The associates of the meanings could be nouns, verbs, or adjectives (mean frequency = 151.6, mean number of letters = 5.8, mean association strength = 0.16). The meanings and their associates can be found in Appendix 5.

A further 34 nouns were selected from the free association corpus to be used in a real word prime condition in the semantic decision task, acting as a comparison with the novel word prime condition. These primes were chosen based on the same criteria as the novel word meanings, except that these items were of lower frequency (mean frequency = 9.0, mean frequency of associates = 66.0, mean number of letters = 5.74, mean number of letters in associates = 6.02). The reason for choosing low frequency words for this control condition was to try to better match them with the novel words which naturally would have extremely low frequencies due to being recently learned. As before, three strong associates were selected for each prime (mean association strength = 0.16). Care was taken in the selection of all items to ensure that there was no overlap: no word was repeated across the primes or the associates. These stimuli are presented in Appendix 6.

Both the training and testing phases included a sentence plausibility judgement task which required generation of sentences in which the novel words could be used. For these tasks, five unique sentences were generated for each novel word, in which the novel word meaning fitted the context of the sentence (e.g., *"the girl was woken up by the paws of her hungry feckton"*, where feckton is a type of cat). The aim was to make these sentences highly constraining so that there would be minimal ambiguity as to whether the novel word in the sentence was used appropriately or not. Four sentences were used in the training task, and one in the test task. One of the four training sentences was combined with a different novel word which made no sense in the context of the sentence (e.g., *"the girl was woken up by the paws of her hungry glain"*, where *glain* is a type of book). These sentences were used in the incorrect usage condition during training (see Procedure). For counterbalancing purposes, the word-sentence pairs used in the test task were randomly divided into two lists. Half of the participants saw one list in the correct usage condition where the novel word was presented with its matching sentence, and the other list in the incorrect usage condition where the novel words were randomly

paired with incorrect sentences. For the other half of participants the list assignment was switched. The sentences used in testing are included in Appendix 7.

Finally, spoken versions of all the novel words as well as the real word primes were recorded by a native English speaker, using the same recording equipment and procedure as in Experiment 3. These auditory stimuli were used in the shadowing task only.

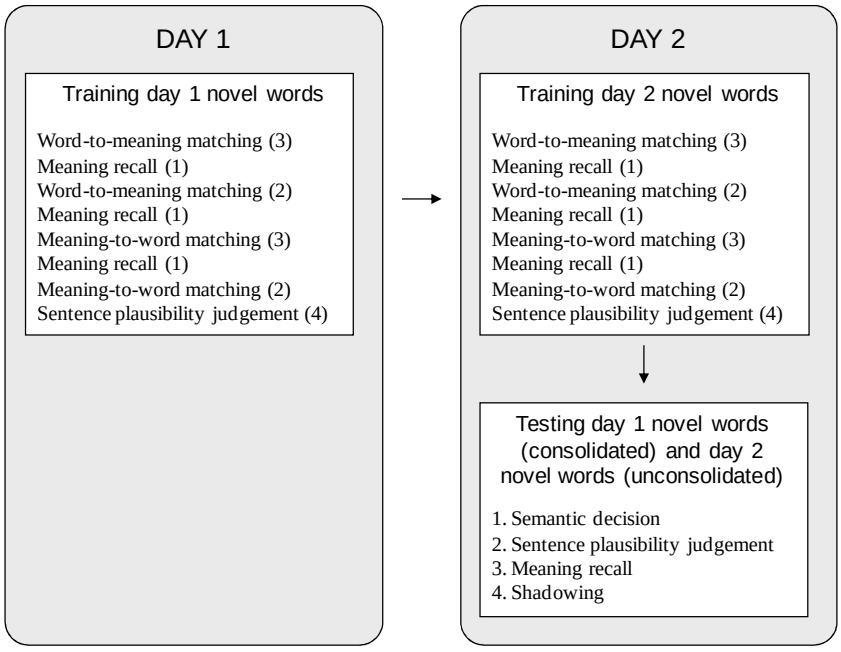
Design

The full set of 102 novel words was divided into three lists of 34 words to be used in the three time of training conditions ("consolidated", "unconsolidated", and untrained). "Consolidated" novel words were words which had been learned on the previous day, while "unconsolidated" novel words were learned just before testing. Note that these terms were chosen for convenience, and do not imply an assumption that the words would actually undergo consolidation within the time scale used in this particular experiment. Whether they do is indeed the question addressed in the experiment. Untrained novel words were stimuli which were not included in training; they acted as a nonword control. The words in the three lists were matched in number of letters (6.4, 6.4, and 6.5). Similar sounding words were avoided within a list, to avoid confusability between words. All three lists were used in all three conditions across all participants. The 68 meanings were divided into two lists of 34 meanings for use in the two time of training conditions (consolidated and unconsolidated). The two lists were matched on frequency (72.3 and 75.0, mean frequency of associates = 150.0 and 152.8), mean association strength (both 0.16), number of letters (4.74 and 4.85), and number of letters in associates (both lists 5.77). Across all participants, both meaning lists occurred in consolidated and unconsolidated conditions an equal number of times. Since the novel words had no existing meaning and did not resemble existing words (Deacon et al., 2004), there was no danger of a given novel word being associated with an assigned meaning prior to training.

Procedure

Training. Figure 12 shows the timing of the training and test sessions, and the tasks used in the sessions. Participants arrived in the lab on day 1 for their first training session, the purpose of which was to train them on the first set of 34 novel

words (consolidated novel words). No testing took place on day 1. They returned on day 2, and carried out the second training session with the second list of novel words (unconsolidated novel words). The testing session immediately followed this second training session. Training of this number of words takes a fairly long time (on average 60-90 minutes), so four different training tasks were used within each training session, to help maintain participants' attention. These tasks consisted of word-to-meaning matching (five exposures to each word), meaning-to-word matching (five exposures), meaning recall (three exposures), and sentence plausibility judgement (four exposures). The training started with three blocks of word-to-meaning matching, with each novel word occurring once in each block. Participants then carried out one block of meaning recall with one exposure to each novel word, followed by two more blocks of word-to-meaning matching. At this point a second block of meaning recall was carried out, followed by three blocks of meaning-to-word matching. A last block of meaning recall was carried out before completing two more blocks of meaning-to-word matching. The final task of the session was sentence plausibility judgement, which included four blocks. Hence the total number of exposures to novel words in this training regime was 17. The meaning recall task was interleaved with the other tasks as outlined above for two reasons. Firstly, it allowed the tracking of learning as a function of amount of



Note. Numbers in parentheses show the number of exposures to each novel word in each training task.

Figure 12. Schematic showing the timing of training and tests.

training. Secondly, it allowed the participants to identify words which they had not yet learned, and to focus on these words as training progressed.

I will now describe the procedure of each training task in detail. In the word-to-meaning matching task a novel word was presented in the middle of the top half of the computer screen, and two meaning alternatives were shown on the left and right sides of the bottom half. The participant was asked to indicate which of the two meaning alternatives was the correct one for the given novel word. The meaning was always presented in the form of "*is a type of X*" where X refers to one of the meanings described above (e.g., *cat*). On each trial, one of the options was correct, and the incorrect option was randomly picked from the pool of meanings used in the current session by the experimental software. The allocation of the correct option to the left or right side was also randomly assigned for each trial. After a response was made by pressing a key labelled "Left" or "Right" on a button box, the incorrect option disappeared from the screen, and the correct one remained on screen for 1500 ms and was followed by a new trial. Unlimited time was given to make the response. Participants were asked to pay close attention to the correct meaning remaining on the screen, in order to start learning the meanings at the beginning of the training. A new random presentation order was used in each of the five blocks.

The meaning-to-word matching task was identical to the previously described task, except that in this task a meaning was presented in the top half of the screen, and two novel word alternatives were presented on the bottom half. Participants were asked to pick the correct novel word for the given meaning. The order of presentation in the five blocks of this task too was randomised.

In the meaning recall task a novel word was presented on the screen, and the task was to type in the meaning of the word by using a standard keyboard. Unlimited time was given for responding. No accuracy feedback was given, but the correct answer was always presented after the response was completed. Order of trials in each block was randomised for each participant.

The fourth training task was the sentence plausibility judgement task. Here a sentence was presented on the screen, using one of the novel words. The participant's task was to indicate whether the novel word fit in the context (correct usage) of the sentence in terms of its meaning or not (incorrect usage). Unlimited time was given to make a response. A response was followed by a feedback screen providing accuracy feedback, the novel word, and its meaning. Each novel word was

presented four times in this task, three times in the correct usage condition, and once in the incorrect usage condition, with a new sentence in each presentation. This imbalance was intentionally built into the design to minimise the presentation of novel words in an incorrect context to avoid interference during learning. The presentation order of the sentences was randomised.

Testing. After completing the training session on day 2, participants carried out the testing session. Participants were offered a chance to take a break between training and testing, but no enforced break was implemented. The order of the testing tasks was fixed (see Table 2). The first testing task was semantic decision. A trial in this task started with the presentation of a fixation cross in the middle of the screen for 500 ms. This was replaced by the prime word, presented for 200 ms, and followed by the target word for 200 ms, with an SOA of 700 ms (ITI 500 ms). The participant was given 2000 ms from the onset of the target to respond. The task was to indicate whether the two words were semantically related or not, by pressing a key labelled “Yes” or “No” on a Cedrus button box. To encourage fast responding, the RT was displayed on screen after a response was made. Accuracy feedback was only given half way through the task in connection with a rest break, in the form of percentage of correct responses made so far. The reason for not giving accuracy feedback after each trial was to avoid any further explicit learning during testing. After inspecting the RT feedback for as long as they wished, participants could initiate a new trial with a button press.

After ten practice trials (using stimuli not seen in the experimental trials), participants did two versions of the semantic decision task. In the novel word version the prime was one of the 68 trained novel words, and the target was a real word semantically related or unrelated to the meaning of the novel word. In the real word version the prime was one of 34 real word primes, followed by a related or an unrelated real word target. The two versions were always done sequentially, with the order counterbalanced across participants.

Due to the limited number of trained novel words acting as primes, and in order to increase the amount of data in this task, it was necessary to present the same prime more than once. The task (both real and novel word versions) was divided into three blocks. Table 2 shows how the primes and targets were distributed over the three blocks. Within each block, each prime occurred twice, once with a semantically related target, and once with an unrelated target. Hence each prime was

Table 2. Sequence of tasks on day 2 of Experiment 4.

Training session: learn novel words and meanings in 4 training tasks		
feckton is a type of cat		
glain is a type of book		
.		
.		
.		
Test 1: semantic decision (prime – target pairs shown below)		
Block 1	Block 2	Block 3
feckton – mouse *(related)*	feckton – kitten *(related)*	feckton- dog *(related)*
feckton – study *(unrelated)*	feckton – school *(unrelated)*	feckton – read *(unrelated)*
glain – read *(related)*	glain – study *(related)*	glain – school *(related)*
glain – kitten *(unrelated)*	glain – dog *(unrelated)*	glain – mouse *(unrelated)*
.	.	.
.	.	.
.	.	.
Test 2: sentence plausibility		
The woman liked to listen to the purring of her feckton *(correct usage)*		
The businessman kept his suits neatly in his glain *(incorrect usage)*		
.		
.		
.		
Test 3: meaning recall		
feckton		
glain		
.		
.		
.		
Test 4: shadowing (auditory presentation)		
/feckton/		
/glain/		
.		
.		
.		

Note. On day 1 only the training session was carried out. Different sets of novel words were learned on the two days. In semantic decision, real word prime condition is not shown in the table.

seen six times in total in this task. Recall that for each prime three related targets were chosen. A different related target was presented in each of the three blocks. The unrelated prime-target pairs were created by pseudorandomly combining a prime with the related target of another prime (while making sure that the resulting pairs were indeed semantically unrelated). So during the task each target occurred twice, once in the related condition and once in the unrelated condition. However, as demonstrated in Table 2, no target ever occurred twice within the same block.

The order of trials in semantic decision was pseudorandomised using Mix (Van Casteren & Davis, 2006). The randomisation constraints stipulated that there had to be at least 15 trials separating the repetition of any given prime or target. Furthermore, only a maximum of three consecutive trials from the same time of

training or relatedness condition were allowed. A new pseudorandomised order was generated for each participant. Half of the participants responded to the related condition with their right hand and unrelated condition with the left hand, the key assignment was swapped for the other half.

The second testing task was sentence plausibility. Only three practice trials were included as the main gist of the task was already familiar to the participants from the training. The only procedural difference in the testing stage was that the sentence was initially presented without the final, novel word. Instead, the final word was marked by a row of four Xs. The task was to read the sentence carefully (there was no time limit for this), and to press a key on the button box to reveal the final word. Timing started from this key press, and participants were required to decide as quickly and as accurately as possible whether the final word was appropriate in the context of the sentence by pressing a key (labelled "Yes" or "No") on the button box. This method controlled for differing reading speeds across participants. Half of the trials were correct usage trials where the novel word matched semantically with the sentence, and half were incorrect usage trials where the novel word was a semantic mismatch. The order of trials was randomised for each participant by the presentation software. No repetition of sentences or novel words was used here. RT feedback was provided in the same way as in the semantic decision, without accuracy feedback. Half of the participants responded to the match condition with their right hand and mismatch condition with the left hand, the key assignment was swapped for the other half.

The third task was meaning recall. This was identical to the recall task used in training, included all 68 novel words presented in random order, and required participants to type in the meaning of a given novel word. No time constraint was set for completing this task.

The final test task was shadowing, the purpose of which was to replicate the shadowing consolidation effect shown by Davis et al. (2009). A shadowing trial began with the presentation of a warning signal (the word READY) on screen for 500 ms. Once the warning disappeared, a spoken word was played through a set of headphones (Beyerdynamic DT 294). These stimuli included the 68 trained novel words, 34 untrained novel words, and 34 real words (these were the same real words that were used as real word primes in the semantic decision task). The participant's task was to repeat the spoken word as quickly and as accurately as possible. After

3000 ms from the onset of the spoken word had elapsed, a new trial was initiated. Stimulus order was newly randomised for each participant, and stimuli were delivered by DMDX (Forster & Forster, 2003), which also recorded and timed the vocal responses. Five practice trials (two real words, three nonwords) were completed before starting the experimental trials. The experimenter remained in the room during practice to ensure participants were speaking loud enough for responses to be recorded. All participants in the test stage were run on a Windows XP PC using a 17" Iiyama Vision Master Pro 454 monitor with a high refresh rate to maximise timing accuracy of visual stimulus delivery.

Participants

Twenty-four University of York students took part in the experiment. All participants were native English speakers with no reported language disorders (11 male, one left-handed, mean age = 20.4, range = 18-23). Participants were paid or received course credit. To provide an extra incentive for both learning and performance in the test tasks, the most accurate and fastest 50% of the participants were entered into a prize draw for a £10 gift certificate.

4.2.2 Results

Training data

Performance in the meaning recall training task was analysed to see if novel word meanings on both training days were learned equally well and because this task is likely to give a more accurate reflection of novel word knowledge than the meaning-to-recall matching data. Figure 13 shows the proportion of correct responses in the three blocks of this task, which were interleaved with the other training tasks. A mixed-effects logistic regression model with subjects and items as random factors, and time of training (consolidated words on day 1, and unconsolidated words on day 2) and training block (three blocks) as fixed factors was fitted. Subject-specific slopes for the effect of time of training significantly improved the fit of the model. No significant interaction contrasts were found, so the interaction was dropped. The simplified model showed significantly better meaning recall on day 2 than day 1 ($b = 0.525$, $z = 3.31$, $p = .001$), and that performance

improved from block 1 to block 2 (b = 1.988, z = 20.30, p < .001), from block 1 to block 3 (b = 3.524, z = 25.19, p < .001), and from block 2 to block 3 (b = 1.534, z = 10.92, p < .001). In the sentence plausibility judgement task very high accuracy rates were seen on both training days (proportion of correct responses was 0.979 on day 1 and 0.977 on day 2). The difference between the two was non-significant when tested with a mixed-effects logistic regression model with subjects and items as random factors, and time of training as a fixed factor.

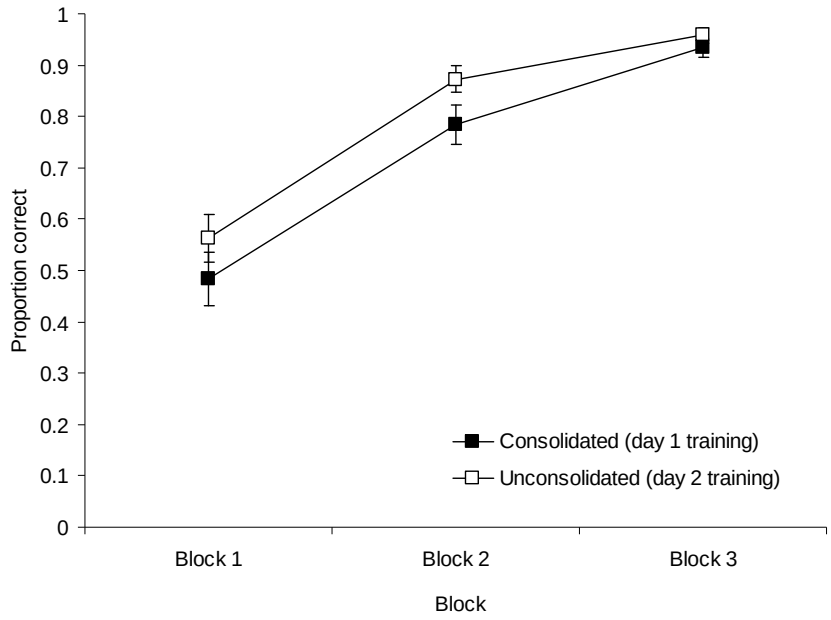


Figure 13. Accuracy rates in the meaning recall training task in each training block. Error bars represent standard error of the means.

Test data

Semantic decision with novel word primes. Reaction time data in the semantic decision task were analysed first. Only correct responses were included in the analysis. As recommended by Baayen (2008), the RT data were log transformed to better satisfy the assumption of normality and the data were trimmed by removing extremely fast or slow responses (RTs faster than 5 log-ms [148 ms] and slower than 7.3 log-ms [1480 ms]) prior to the analysis. A mixed-effects linear model with subjects and items as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated), and semantic relatedness of the prime and target (related vs. unrelated) as fixed variables was built. Subject-specific slopes for relatedness and time of testing significantly improved the fit of the model, as did

item-specific slopes for time of testing. Figure 14 (left panel) shows the RT data for both time of testing and relatedness conditions. There was no interaction between the two variables, so it was dropped. The simplified model showed significantly faster semantic decisions to the related word pairs than the unrelated word pairs ($b = 0.089$, $t = 4.49$, $p < .001$). The overall RT difference between the consolidated and unconsolidated conditions was non-significant. The effect of relatedness was also analysed for each time of training condition individually. The effect was significant both in the consolidated ($b = 0.084$, $t = 4.34$, $p < .001$) and unconsolidated conditions ($b = 0.098$, $t = 5.09$, $p < .001$). The effect of time of training was not significant in the related or the unrelated trials.

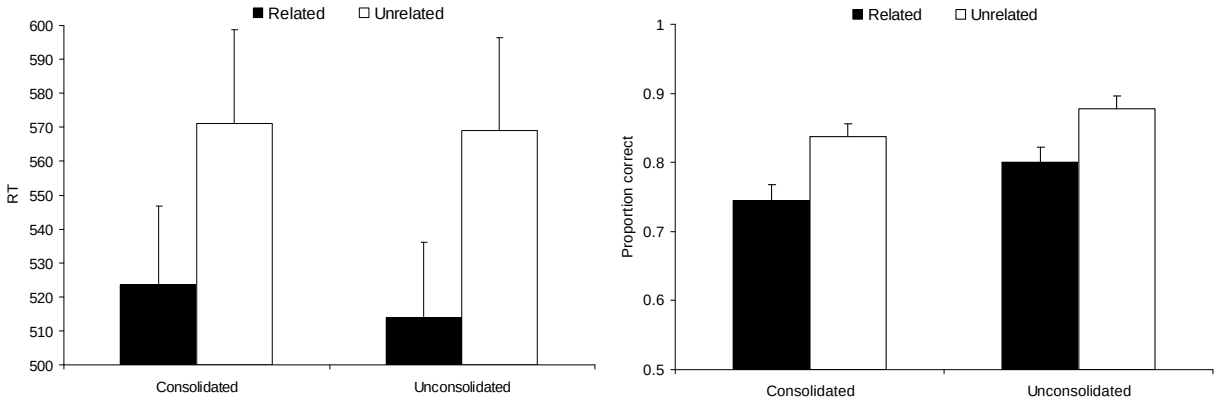


Figure 14. RTs (left panel) and accuracy rates (right panel) in the semantic decision task with novel word primes. Error bars represent standard error of the means.

Accuracy rates were analysed next (Figure 14, right panel). A mixed effects logistic regression model was used, with subject- and item-specific slopes for the relatedness condition. As usual, subjects and items were entered as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated), and semantic relatedness of the prime and target (related vs. unrelated) as fixed variables. No significant interaction was found. Related word pairs attracted fewer accurate responses than unrelated pairs ($b = -0.592$, $z = -2.88$, $p = .004$), and consolidated novel word trials had lower accuracy rates than unconsolidated trials ($b = -0.368$, $z = -6.53$, $p < .001$). Looking at the effect of relatedness at the two levels of the time of training condition, the effect was significant in the consolidated condition ($b = 0.593$, $z = 2.79$, $p = .005$) and in the unconsolidated condition ($b = 0.588$,

z = 2.74, p = .006). The effect of time of training was significant in the related condition (b = 0.369, z = 4.99, p < .001), and in the unrelated condition (b = 0.366, z = 4.20, p < .001).

As analysed above, the data collapsed across the three blocks showed a relatedness effect for both consolidated and unconsolidated novel words, and did not show overall RT differences between the consolidated and unconsolidated novel word trials (although a difference in accuracy was found). To see if this RT pattern was true at early and late stages of the task, the same data were analysed including experimental block as an additional fixed variable in the model described above. Figure 15 shows the RT data broken down by block. A three-way interaction contrast showed that the effect of relatedness changed over blocks significantly more in unconsolidated than consolidated items (b = -0.074, t = -2.64, p = .006[†]): in consolidated items the difference between related and unrelated trials remained stable over blocks, but in unconsolidated items it was larger in block 1 than block 3. The effect of relatedness in consolidated novel words was significant in all three blocks (block 1: b = 0.095, t = 4.02, p < .001, block 2: 0.076, t = 3.24, p < .001, block 3: 0.091, t = 3.89, p < .001). In the unconsolidated condition also the relatedness effect was significant in all blocks (block 1: b = 0.140, t = 6.05, p < .001, block 2: 0.090, t = 3.85, p < .001, block 3: b = 0.061, t = 2.64, p < .001).

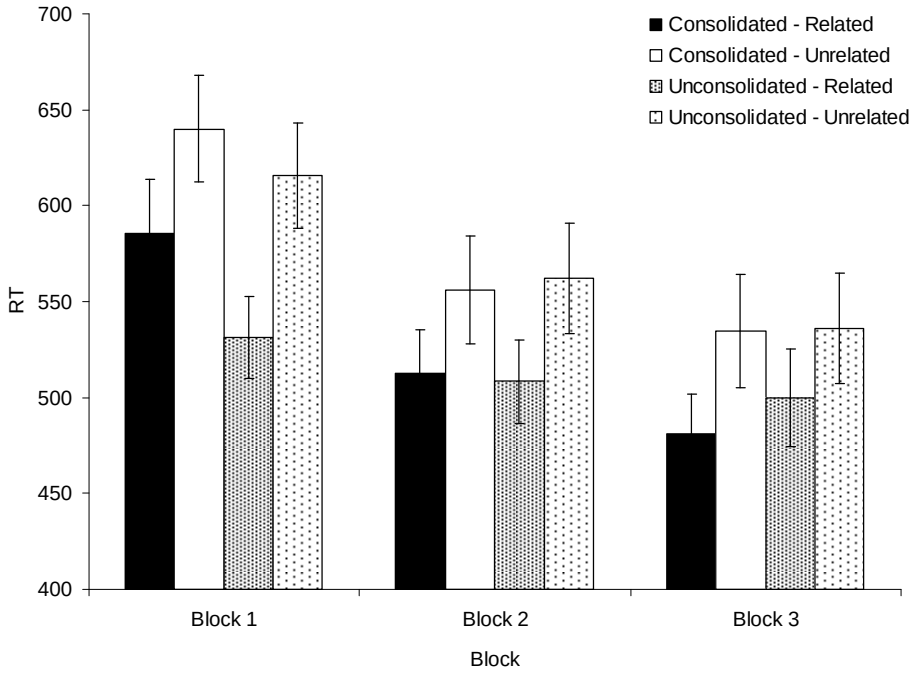


Figure 15. Semantic decision RTs in each block with novel word primes. Error bars represent standard error of the means.

Visual inspection of Figure 15 suggests that while semantic decisions to unconsolidated prime trials are faster in the first block, the pattern reverses in the final block. Contrasts involving block and time of testing showed that in related trials responses to unconsolidated words were significantly faster than to consolidated words in block 1 ($b = -0.086$, $t = -5.99$, $p < .001$), the two were equally fast in block 2, but in block 3 responses to unconsolidated words were now significantly slower ($b = 0.035$, $t = 2.43$, $p = .02^{\dagger}$). In unrelated trials in block 1 unconsolidated words were responded to significantly faster than consolidated novel words ($b = -0.041$, $t = -2.89$, $p < .001$), but the effect failed to reach significance in the two later blocks.

Finally, the effect of block was evaluated in each relatedness and time of training condition. Responses to consolidated novel words in the related condition became significantly faster from block 1 to block 2 ($b = -0.128$, $t = -8.75$, $p < .001$), from block 1 to block 3 ($b = -0.187$, $t = -12.70$, $p < .001$), and from block 2 to block 3 ($b = -0.059$, $t = -4.04$, $p < .001$). The same contrasts in the unconsolidated condition showed faster responses from block 1 to block 2 ($b = -0.047$, $t = -3.35$, $p < .001$), from block 1 to block 3 ($b = -0.066$, $t = -4.73$, $p < .001$), but not from block 2 to block 3. In the unrelated condition responses to consolidated novel words again became faster from block 1 to block 2 ($b = -0.148$, $t = -10.52$, $p < .001$), from block 1 to block 3 ($b = -0.192$, $t = -13.75$, $p < .001$), and from block 2 to block 3

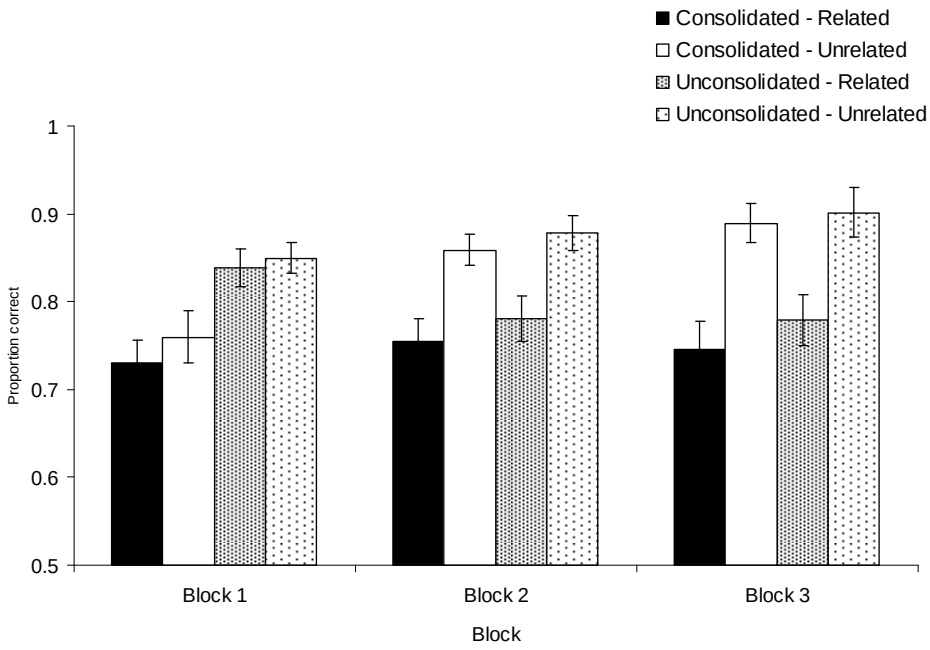


Figure 16. Accuracy rates in semantic decision in each block. Error bars represent standard error of the means.

(b = -0.044, t = -3.26, p = .001). In this condition responses to unconsolidated novel words became faster from block 1 to block 2 (b = -0.977, t = -7.17, p < .001), from block 1 to block 3 (b = -0.146, t = -10.76, p < .001), and from block 2 to block 3 (b = -0.048, t = -3.58, p < .001).

Figure 16 shows accuracy rates in semantic decision, broken down by block. Block was again added as a fixed factor to the logistic regression model described earlier. No contrasts involving three-way interactions reached significance. In the simplified model contrasts involving the interaction between block and relatedness showed that the accuracy difference between related and unrelated trials was significantly larger in block 2 compared to block 1 (b = 0.608, z = 4.45, p < .001), and in block 3 compared to block 1 (b = 0.979, z = 6.83, p < .001). Looking at the effect of relatedness on each block, there was no significant difference between related and unrelated trials in the first block in either consolidated or unconsolidated conditions. In the second block the relatedness effect reached significance in both consolidation conditions (consolidated: b = 0.737, z = 3.21, p = .001, unconsolidated: b = 0.689, z = 2.98, p = .002). The same was true in the third block (consolidated: b = 1.109, z = 4.76, p < .001, unconsolidated: b = 1.059, z = 4.51, p < .001). Interaction contrasts involving time of training and block suggested that time of training had a significantly larger effect in block 1 than either in block 2 (b = -0.536, z = -3.93, p < .001) or block 3 (b = -0.511, z = -3.65, p < .001). Averaged over relatedness (in the absence of a three-way interaction), consolidated novel words attracted more errors than unconsolidated novel words in block 1 (b = 0.704, z = 7.43, p < .001). The effect was non-significant in block 2, but approached significance in block 3 (b = 0.193, z = 1.88, p = .06†). Finally, the effect of block was evaluated in each time of training and relatedness condition. In related trials, the consolidated novel word condition was unaffected by block. In the unconsolidated condition however number of accurate responses in the related condition declined both in block 2 and block 3 compared to block 1 (b = -0.391, z = -3.29, p = .001, b = -0.402, z = -3.35, p < .001 respectively). No change from block 2 and block 3 was seen. In unrelated trials, accuracy in consolidated novel words improved from block 1 to block 2 (b = 0.761, z = 6.39, p < .001) and block 3 (b = 1.098, z = 8.61, p < .001), and from block 2 to block 3 (b = 0.336, z = 2.49, p = .01†). In the unconsolidated condition accuracy increased from block 1 to block 3 (b = 0.580, z = 4.28, p < .001) and from block 2 to block 3 (b = 0.359, z = 2.57,

$p = .01^{\dagger}$), but not from block 1 to block 2.

To summarise the main findings, when RTs in the semantic decision task were analysed averaged over the three blocks, there was no evidence of a consolidation benefit: RTs to trials with consolidated novel word primes were not significantly different from unconsolidated trials, and in the accuracy rates lower accuracy was found for consolidated trials. However, when the data were broken down by block, allowing an examination of RTs in the early and late stages of the task (block 1 vs. block 3), some preliminary evidence for a consolidation benefit was found: in the third and final block RTs to consolidated novel word trials were faster than unconsolidated trials, although this effect was significant only in the related prime-target condition. No such effect was found in accuracy rates though.

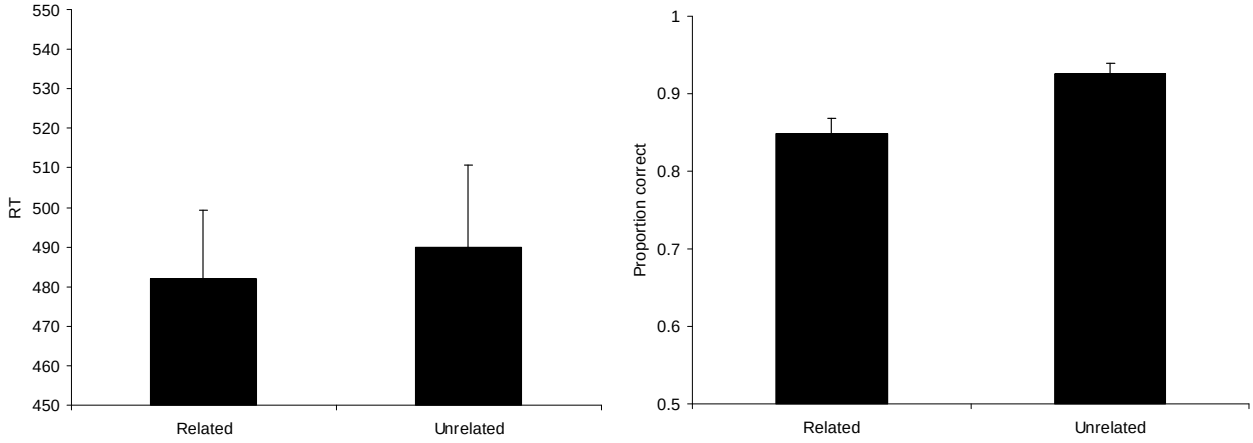


Figure 17. RTs (left panel) and accuracy rates (right panel) in the semantic decision task with real word primes. Error bars represent standard error of the means.

Semantic decision with real word primes. Figure 17 shows RTs to the semantic decision task when both the prime and target were real, familiar words. Averaged across the three blocks (left panel in Figure 17), no significant difference was found between related and unrelated trials in a mixed-effects linear model with subjects and items as random factors, and relatedness (related vs. unrelated) as the fixed factor (with subject- and item-specific slopes for relatedness). Accuracy rates are shown in the right panel of the figure. Here a mixed effects logistic regression model with the same structure as above showed a significant difference between the relatedness conditions: accuracy rates were significantly higher in the unrelated condition ($b = 0.830$, $z = 3.15$, $p = .002$).

To examine the effect across the three blocks of the task (Figure 18), block was added as a fixed factor (in this model subject-specific slopes for the effect of block increased goodness of fit). Interaction contrasts involving relatedness and block showed that the relatedness effect in block 2 was significantly smaller than in block 1 (b = -0.034, t = -2.09, p = .04†) and in block 3 (b = -0.042, t = -2.60, p = .01†). Hence relatedness was evaluated at each block separately. The effect was marginally significant in block 1 (b = 0.022, t = 1.98, p = .05†), non-significant in block 2, but significant in block 3 (b = 0.031, t = 2.68, p = .009). Looking at the effect of block next, RTs to related trials did not differ significantly across blocks. The same was true of the unrelated trials.

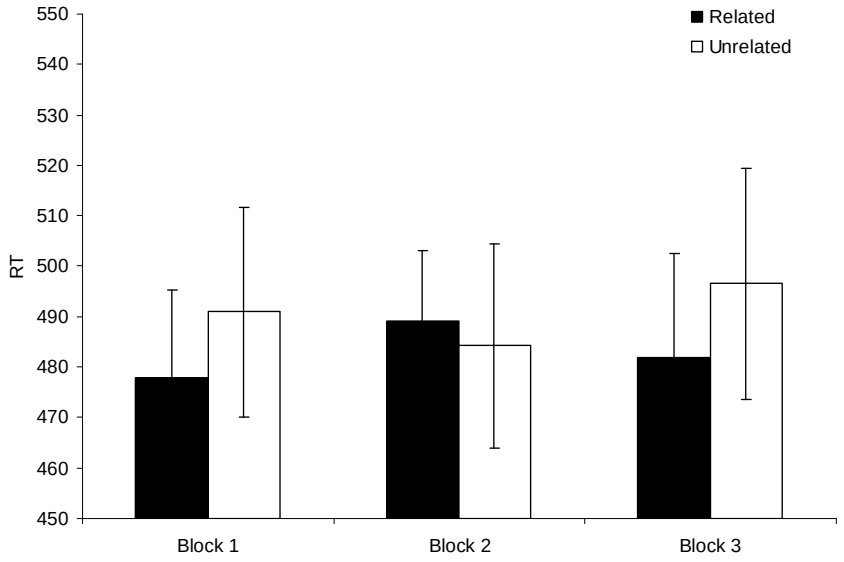


Figure 18. Semantic decision RTs in each block with familiar word primes. Error bars represent standard error of the means.

Figure 19 shows the accuracy rates in each block. To analyse these data block was added as a fixed factor to the model used in the accuracy analysis above. Interaction contrasts involving block and relatedness showed that the relatedness effect was significantly larger in block 2 compared to block 1 (b = 0.781, z = 3.05, p = .002). No difference in the magnitude of the relatedness effect was found in the other contrasts. The effect of relatedness was evaluated in each block next. No effect was found in block 1, but it did reach significance in block 2 (b = 1.194, z = 3.92,

$p < .001$) and block 3 ($b = 0.840$, $z = 2.83$, $p = .005^{\dagger}$). Looking at the effect of block in each condition, accuracy rate declined in the related condition from block 1 to block 2 ($b = -0.500$, $z = -3.25$, $p = .001$), and from block 1 to block 3 ($b = -0.438$, $z = -2.82$, $p = .005^{\dagger}$). No significant decline was seen between block 2 and block 3. In the unrelated condition, there was no change in accuracy rate across the blocks.

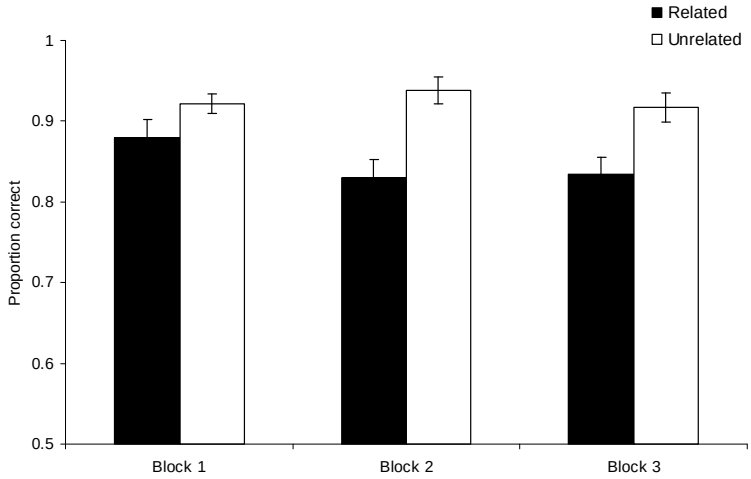


Figure 19. Accuracy rates in semantic decision in each block using familiar primes. Error bars represent standard error of the means.

Sentence plausibility judgement. RTs in the sentence plausibility judgement task (Figure 20, left panel) were analysed using a mixed-effects linear model with subjects and items as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated) and the semantic compatibility of the novel word in the sentence context (match vs. mismatch) as fixed variables. Subject-specific slopes for sentence-word compatibility increased the goodness of fit. The same data trimming criteria were used here as in the semantic decision task. Visual inspection of RTs for each participant suggested that as the task progressed, some participants became slower in responding, while other became faster. Such variability can be accounted for in the model by adding subject-specific slopes for trial position. This significantly increased the goodness of fit. A significant interaction was found between semantic compatibility and time of testing, suggesting that the effect of compatibility was smaller in unconsolidated items ($b = -0.063$, $t = -2.25$, $p = .01$). The effect of compatibility was significant in both time of testing conditions (consolidated: $b = 0.184$, $t = 6.73$, $p < .001$, unconsolidated: $b = 0.121$, $t = 4.44$, $p < .001$). Next the effect of time of testing was evaluated in each compatibility condition.

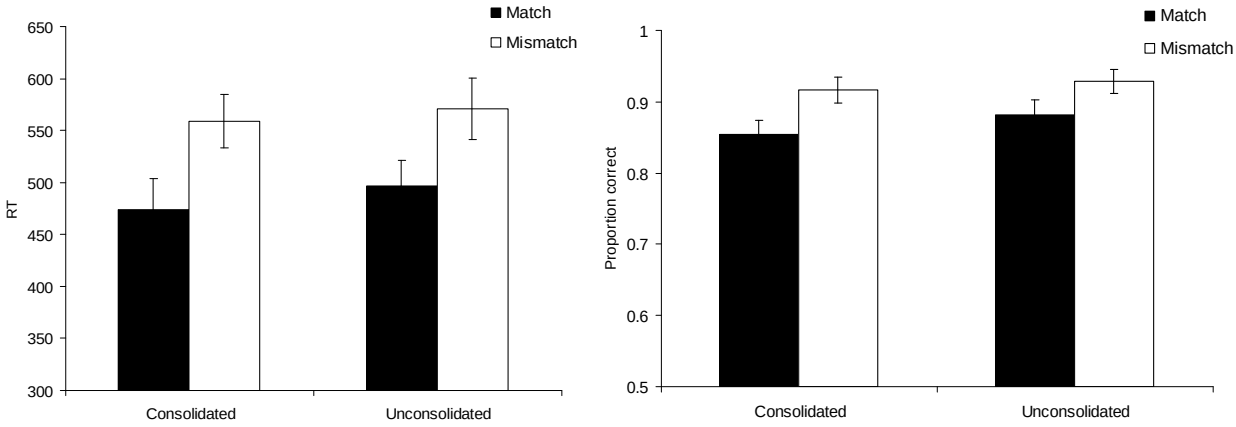


Figure 20. RTs (left panel) and accuracy rates (right panel) in the sentence plausibility judgement task. Error bars represent standard error of the means.

In the match condition consolidated words were responded to significantly faster than unconsolidated words ($b = 0.070$, $t = 3.56$, $p < .001$). In the mismatch condition this difference was non-significant. It seems then that in this task there was an RT advantage for semantically compatible novel words. Interestingly, this advantage was larger for consolidated novel words, possibly suggesting that the meaning of these words was accessed faster than the meaning of unconsolidated novel words.

Accuracy rates in the sentence plausibility task are shown in the right panel of Figure 20. A mixed effects logistic regression model with subjects and items as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated) and semantic compatibility of the novel word and the sentence (match vs. mismatch) as fixed variables showed no interaction between the two variables. Averaged over the time of training conditions mismatch trials had more accurate responses than match trials ($b = 0.668$, $z = 3.48$, $p < .001$). There was no significant difference between consolidated and unconsolidated novel words when averaged across compatibility conditions. Although the lack of an interaction suggests that time of training had no effect on performance, the sentence-word compatibility effect was evaluated for consolidated and unconsolidated novel words separately to ensure the effect was significant in both conditions. This was the case (consolidated: $b = 0.715$, $z = 2.78$, $p = .006$, unconsolidated: $b = 0.616$, $z = 2.23$,

$p = .03^{\dagger}$). Furthermore, looking at match trials only there was no significant difference between consolidated and unconsolidated conditions, and the same was true of mismatch trials.

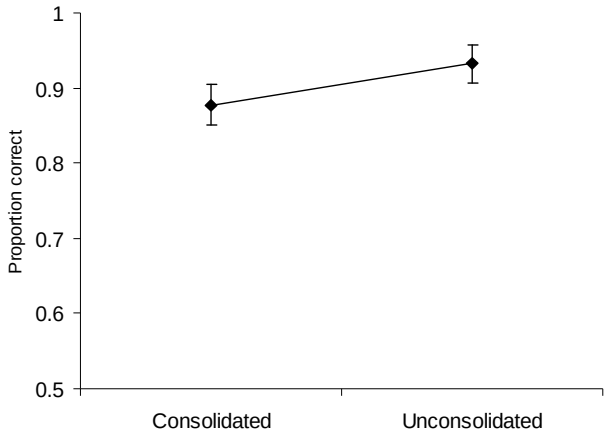


Figure 21. Accuracy rates in the meaning recall test task for consolidated and unconsolidated novel words. Error bars represent standard error of the means.

Meaning recall. Figure 21 shows accuracy in the meaning recall test task. One participant's data were lost in this task due to equipment malfunction. A mixed effects logistic regression model with subjects and items as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated) as the fixed variable showed that participants recalled more unconsolidated word meanings compared to consolidated word meanings ($b = 0.786$, $z = 4.01$, $p < .001$).

Shadowing. To ensure reliable reaction time data in the shadowing task, the voice key trigger points were checked manually using the CheckVocal software (Protopapas, 2007), and corrected when necessary. Repetition latencies and accuracy rates are presented in Figure 22. Latencies to accurate responses (left panel) were analysed using a mixed-effects linear model with subjects and items as random variables, and training condition (consolidated, unconsolidated, untrained, real words) the fixed variable. Subject-specific slopes for the effect of trial increased the goodness of fit of the model. The same data trimming criteria were used here as in the semantic decision task. The analysis showed that consolidated novel words were shadowed faster than unconsolidated ($b = 0.017$, $t = 3.49$, $p < .001$) and untrained novel words ($b = 0.024$, $t = 5.03$, $p < .001$), but did not differ from real words.

Unconsolidated novel words on the other hand did not differ significantly from untrained novel words, but were responded to significantly slower than real words (b = -0.038, t = -2.89, p = .003). Finally, the difference between response latencies to untrained novel words and real words was significant (b = 0.046, t = 3.45, p < .001).

The accuracy rates (Figure 22, right panel) reflected the latency data, as shown by a mixed effects logistic regression model with subjects and items as random variables, and training condition (consolidated, unconsolidated, untrained, real words) as the fixed variable. Accuracy for consolidated novel words was significantly higher than unconsolidated novel words (b = -0.832, z = -2.36, p = .02†) or untrained words (b = -1.039, z = -3.04, p = .002), but not significantly different from real words. Unconsolidated novel words did not differ significantly from untrained novel words, while real words resulted in marginally higher accuracy rates than unconsolidated novel words (b = 0.772, z = 1.86, p = .06†). Real words did have higher accuracy rate than untrained novel words (b = -0.978, z = -2.39, p = .02†).

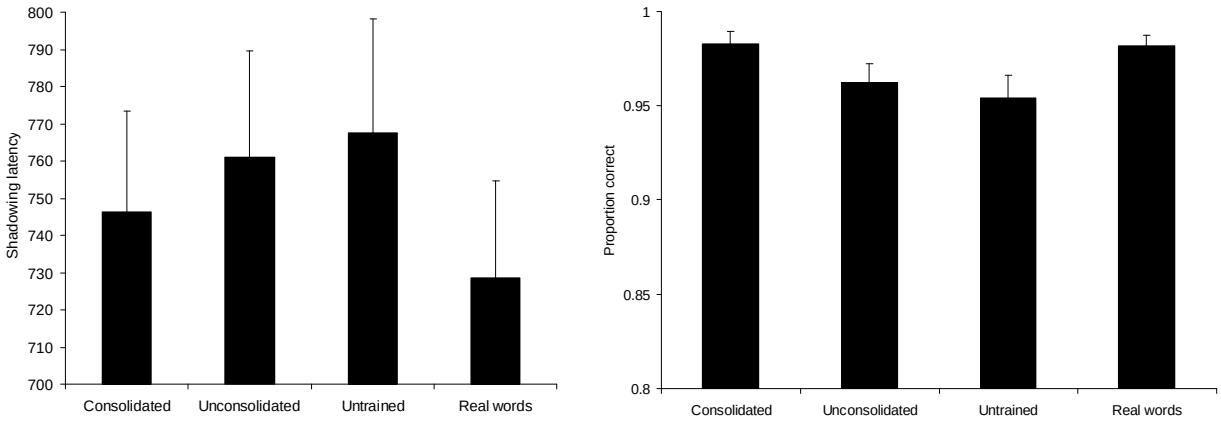


Figure 22. Shadowing latencies (left panel) and accuracy rates (right panel). Error bars represent standard error of the means.

4.2.3 Discussion

This experiment included a number of tasks attempting to measure the speed and accuracy of access to novel word meanings. A semantic relatedness effect was found in the semantic decision task using real word primes and targets (although it was only seen in two of the three blocks of the task), whereby faster responses were made in trials where the prime and the target were semantically related. This may reflect a priming effect where the recognition of the prime facilitates recognition of a

semantically associated target, leading to faster RTs in related trials compared to unrelated trials. If novel word meanings have been learned equally well, a similar finding should be made when the prime is a trained novel word. The high accuracy rates in the meaning recall, semantic plausibility, and semantic decision tasks show that participants were able to explicitly learn the meanings of the novel words well. The current data from the semantic decision task replicate earlier reports of faster semantic decisions being made in related trials compared to unrelated trials when the prime is a newly learned word (Figure 14), supporting the notion that the novel word meanings were available even in a speeded task requiring fast semantic access. This effect was found both for consolidated novel word primes (words learned a day before testing) and for unconsolidated novel word primes (words learned immediately before testing). This is the first experiment to my knowledge which in this task compares novel words that either had or did not have a chance to consolidate prior to testing. However, when averaged across blocks, the relatedness effect did not seem to be affected by the time of training manipulation: the advantage for related trials was equally large in both conditions.

One piece of evidence showing a consolidation advantage in this task comes from looking at the RTs in each of the three blocks in the task separately. While there was initially an RT advantage (and an accuracy advantage) for recently learned, unconsolidated novel word trials, this advantage shrank as the task progressed. In the third block responses to related trials in the consolidated condition were in fact faster than responses to the unconsolidated condition (although no difference was found in unrelated trials). This reversal in the time of training effect during the course of the task is potentially important. The explicit meaning recall task showed that participants could recall more meanings of the unconsolidated novel words compared to consolidated words. This suggests that when participants are explicitly trying to access the meanings of the novel words, there is an advantage for recently learned materials. It is possible that this is what they were also doing in the first block of the semantic decision task. However, as they entered the second and the third blocks, response times overall became faster, possibly indicating a shift from effortful, explicit retrieval of the meanings to more online access. Importantly, no such speeding up was seen in the real word condition, presumably as effortless access is always available for real words. The accuracy data showed a similar pattern, with more errors made to consolidated novel words in the first block, but the

difference attenuating in the last two blocks. This supports the idea of initial effortful retrieval of novel word meanings learned a day before, which however becomes more facilitated with further exposures and practice during the task.

Another prominent trend in the semantic decision data was the lower accuracy rate to related trials compared to unrelated trials. The pattern was seen both in the real word and novel word prime conditions, and may reflect a bias in responding, in that participants may have favoured classifying the word pairs as unrelated when they were unsure about the correct response. This would lead to high accuracy rates in unrelated trials and lower accuracy in related trials, as was seen both in novel word and real word prime conditions. This might mean that the faster responses to related trials were at least partially due to a speed-accuracy trade off. However, this is unlikely to be the sole explanation considering that the accuracy difference was not seen in the first block of either the real word or the novel word prime conditions, and yet the relatedness effect was present in the first block as well.

The sentence plausibility judgement task provided potential evidence for a semantic consolidation process. Participants were faster to decide whether a novel word was used in a semantically appropriate sentence when that novel word fit the semantic context of the sentence. Importantly, the word-sentence compatibility effect was larger in consolidated novel words than unconsolidated novel words. This appears to have been caused by faster responses to compatible words when they were consolidated rather than unconsolidated. Again, this suggests that participants were able to access the meaning of the consolidated novel words faster, and proceed to make a semantic judgement about them faster. No accuracy difference was found between the two time of training conditions though. It is important to note however that as this task was self-paced, it is possible that the consolidation effect is caused not by speeded access to meaning, but access to orthography. The long delay between viewing the sentence and revealing the novel word may have allowed the generation of expectancy, for example the participant may have deduced based on the beginning of the sentence that the final word will be *feckton*. When the final word eventually is revealed, all that remains for the participant to do is to match the revealed word with the expected word. Such a matching process may rely more on orthographic processes than semantic ones. This alternative explanation will be outlined in more detail in the General Discussion to the present chapter.

Finally, when access to novel word meaning was measured in an explicit recall task with no time pressure, an advantage for recently learned, unconsolidated novel words was found. The same pattern was found in the meaning recall task in Experiment 1, where higher accuracy was found in participants who were tested immediately after training compared to participants who were tested one day later. This dissociation between a benefit for unconsolidated novel words in a task of explicit meaning recall, and a benefit for consolidated words in tasks requiring speeded meaning access (semantic plausibility and semantic decision) may prove to be important and will be examined further in the next chapter.

The last task in the current experiment was shadowing, a task which primarily measures access to phonological word form representations rather than knowledge of meaning. Here a clear time of training effect was found. Consolidated novel words were shadowed significantly faster than unconsolidated novel words or untrained novel words. Unconsolidated novel words on the other hand were shadowed as slowly as untrained words. Accuracy rates showed exactly the same pattern. Like the lexical competition studies reviewed earlier (e.g., Dumay & Gaskell, 2007), these data suggest that novel word form representations benefit from a period of offline consolidation (at least phonological ones based on the shadowing task data). It is also possible to postulate that a form-based lexical representation was generated for the consolidated novel words, but not for the unconsolidated novel words. This argument relies on the observation that unconsolidated novel words were shadowed as slowly as untrained novel words. When shadowing nonwords (untrained novel words), participants are recreating a novel sequence of phonemes which does not map onto any known lexical representation. The failure to find any difference in shadowing latencies between untrained and unconsolidated words suggests that in neither case was there a lexical representation to facilitate performance (or the emerging lexical representation was too weak to show any benefit). Consolidated novel words on the other hand were shadowed significantly faster (as fast as real words), suggesting that these words had generated a robust lexical representation.

The consolidation effect in shadowing might have a semantic locus too. As people recognise a spoken novel word, the meaning of that word is also activated (e.g., Zwitserlood, 1989). A stronger (consolidated) semantic representation may aid in the recognition and repetition of the novel word, and give rise to faster shadowing

times (shadowing latencies are known to be subject to semantic influence, see for example Slowiaczek [1994] for semantic priming effects in shadowing). Experiment 5 considered the contribution of semantics to the shadowing consolidation effect. If the consolidated novel words were shadowed faster because they activated a semantic representation more strongly than the unconsolidated words, then the shadowing effect should not be observed in novel words for which no meaning was taught, irrespective of time of testing. This hypothesis was tested in Experiment 5.

Another distinctive property of the shadowing task in this experiment was the fact that participants never heard the phonological form of the novel words until they carried out the shadowing task. All training tasks and test tasks were carried out in the visual modality. Hence the phonological form representation that was being probed in the shadowing task had been generated by indirect exposure to phonology via orthographic input. It may be the case that such indirectly generated representations benefit more from consolidation than directly generated representations. To see how this might work, recall that CLS accounts suggest that consolidation occurs as a result of offline reinstatement of the novel memory trace. Training itself offers one way of reinstatement, but requires prolonged exposure to the novel stimulus in the form of repeated training trials. A visual presentation of a novel word is likely to activate a corresponding phonological representation on most trials, but this activation would probably be weaker than direct exposure to a spoken version of the stimulus. At the end of the training the novel phonological trace would be weaker than the orthographic trace, and be in greater need of reinstatement during following offline periods, such as sleep. To see if this is the case, a comparison needs to be made between tasks probing representations generated as a result of direct exposure (a written novel word as a result of written training), and representations generated as a result of indirect exposure (a spoken novel word as a result of written training). In a similar vein, the shadowing data could also be viewed as an example of cross-modal priming. If repeated exposure to the novel word during training gives rise to form-based repetition priming, it may be the case that cross-modal priming requires consolidation to emerge. In that case no effect of consolidation should be seen when tested in the same modality. Experiment 5 was designed to address these last points.

The fixed order of the testing tasks may also have had an impact on the observed patterns of data. For example, semantic decision was always the first of the

testing tasks. This may have contributed to the way the consolidation effect emerged over the course of the task, in particular if the consolidated novel words benefit from a small number of retrieval attempts to heighten their level of activation and to bring it in line with the activation of the recently learned unconsolidated words. Secondly, the fact that the shadowing task was always the last task, may have increased the semantic contribution to the task, as the meaning of the novel words had been accessed repeatedly prior to the shadowing task. Experiment 5 ought to clarify this latter point. In Experiment 6 the meaning recall task was moved to the first position to make sure all novel words had benefitted from meaning retrieval before starting tasks measuring speeded access to meaning. This might remove the effect of block in these types of task.

4.3 Experiment 5

Experiment 5 used an identical shadowing task to Experiment 4, but added a naming (reading aloud) task to evaluate consolidation effects in a task which probes novel word form representations in the same modality they were trained in (orthographic). As mentioned in the preceding discussion, if the indirectly trained representation (phonological) benefits from offline consolidation more than directly exposed representations, a significantly larger difference between consolidated and unconsolidated words should be seen in the shadowing task than in the reading aloud task. Furthermore, to see if these consolidation effects have a meaning or form-based locus, half of the novel words were trained with a meaning, while for the other half only the orthographic form was ever presented. If the locus is indeed semantic rather than form-based, we should see consolidation effects in the meaningful novel words only. A visual cued recall task was also included as a second task to evaluate the accuracy of the novel orthographic representations. Apart from these new tasks and the meaning manipulation the same design and stimuli were used in this experiment and the previous experiment, except for the novel word meanings which were made more realistic, as described below.

4.3.1 Method

Materials

The same 102 novel words and 68 meanings were used as in Experiment 4. The meanings were elaborated for the purposes of this experiment by adding two features to each meaning to generate an object for which no existing word exists. For example, if in Experiment 4 *feckton* was defined as "a type of cat", it was now defined as "a type of cat that has stripes and is bluish-grey". This was done in order to better simulate the kind of word learning that goes on in real life situations. In reality, novel words usually refer to novel objects for which a person does not know a label. By adding new features to the existing meanings from Experiment 4 the aim was to teach a label for a new semantic object rather than teaching a new label for an existing object. The new features were chosen so that they set the object apart from any known object, however without making the new meaning implausible (e.g., the existence of a stripy bluish-grey cat is plausible, but the existence of a cat with six legs would be less so). Again, the plausibility requirement was set to make the learning more realistic. The complete set of elaborated meanings can be found in Appendix 8.

The majority of sentences used in the semantic plausibility judgement task at training were also the same sentences as used in Experiment 4. However, in some cases the sentence context was no longer appropriate for the meaning with the new features, and in these cases the sentence was modified to fit with the elaborated meaning. For example, one of the original sentences to be used with the meaning "a type of baby" was "The couple desperately wanted to have another [a type of baby]". In the current experiment the meaning with new features was "a type of baby that is premature and underweight", making the original sentence implausible. The modified set of testing sentences is found in Appendix 9.

The only task that required completely new materials was the cued recall task. Four different cues were created for each novel word, three to be used in training and one in testing. The first training cue was generated by removing one letter from each novel word (for example *_eckton* from *feckton*). Across all words, all letter positions were used an equal number of times to make sure participants

overall had to keep attending to all parts of the words during training. The missing letter was always replaced with an underscore. The second cue type was created by removing every other letter (*f_c_t_n*), starting with the second letter. The third version was created in the same way, but now the previously removed letters were kept in and the others were removed (*_e_k_o_*), starting with the first letter. The fourth version, to be used in the testing session, was created by removing all but the first, last, and one medial letter (*f_ _ k _ _ n*). Care was taken in choosing the medial letter to make sure that there were no identical fragments referring to two different words, and that each fragment could possibly only accommodate one specific novel word.

Design

The same division of the 102 novel words into three lists of 34 words that was used in Experiment 4 was again used here, for the purposes of dividing the items into consolidated, unconsolidated, and untrained conditions. These lists were further randomly divided into two sub-lists of 17 items. Meaning was taught for words in one sub-list, and no meaning was given for the words in the other sub-list. The 68 meanings were also divided into two lists of 34, as in Experiment 4, to be used in the consolidated and unconsolidated conditions. As before, each meaning list was combined with each novel word list an equal number of times.

Procedure

Training. Participants received their first training session on day 1, during which they were trained on the first set of novel words (consolidated). No testing took place on this day. They returned on the following day and carried out an identical training session on the second set of novel words (unconsolidated). The testing session followed immediately after the day 2 training session. The training was largely similar to that used in Experiment 4, including word-to-meaning matching, meaning-to-word matching, meaning recall, and the sentence plausibility judgement task. These tasks were presented in the same order and the same number of times as in Experiment 4 (i.e., three blocks of word-to-meaning matching, one block of meaning recall, two blocks of word-to-meaning matching, one block of meaning recall, three blocks of meaning-to-word matching, one block of meaning recall, two blocks of meaning-to-word matching, and four blocks of semantic

plausibility judgement), to give a total of 17 exposures to each word. Since half of the novel words were to be trained in the absence of meaning, new meaningless variants of word-to-meaning matching, meaning-to-word matching, and meaning recall were created, and were used only in trials involving the meaningless group of novel words.

In the blocks of word-to-meaning matching task, the word-to-target matching task replaced the word-to-meaning matching trials when a meaningless novel word was presented. This variant presented two target letter options instead of two meaning options. The task was to indicate which of the two letters could be found in the word seen on the screen. Matching the structure of the two tasks closely ensured that both meaningful and meaningless novel words were trained in a similar way, but no meaning was provided for the meaningless items. One of the target letters, randomly allocated to the left or the right, was always correct, and the other randomly chosen from a pool of letters used as targets but not found in the word in question. Target letters were chosen for each word so that across the training session first, medial, and final letter positions were targeted. Also, all target letters occurred an equal number of times as correct targets and as foils so that a response could never be predicted on the basis of the letter alone. In the target-to-word matching (derived from meaning-to-word matching) one target letter was provided with two novel word options. The foil word here was picked from a pool of novel words in the meaningless condition that did not contain the target letter, while ensuring that all novel words appeared as foils an equal number of times. Note that word-to-meaning matching was used on trials involving a meaningful novel word, and word-to-target matching was only used when the novel word to be trained was of the meaningless group. The two word types and their corresponding task variants were randomly intermixed in the presentation order.

The third new training task was cued recall, which was always integrated with meaning recall. A trial in this new cued recall/meaning recall composite task began with the presentation of a cue (e.g., *_eckton*), and the participant was required to type in the complete novel word to which the cue referred to. Once this response had been completed, the complete word was shown on screen, and now the participant was asked to type in the meaning of the word. If it was one of the meaningless novel words, participants typed “no meaning”. Unlimited time was given for all responses. As before, this task was done three times during training. In

the first block the cues were missing one letter (see Materials). In the second and third blocks they were missing every other letter. For half of the words the missing letters started from the first one in the second block (e.g., *f_c_t_n*) and from the second letter in the third block (e.g., *_e_k_o_).* The opposite was true for the other half. The order of trials within blocks was randomised by the software.

One further change was introduced in the meaning recall part of this task. In the first block of meaning recall participants were asked to type in only the object to which the novel words referred to (e.g., *a type of cat*). In the second block they were asked to type in at least one of the features as well (e.g., *a type of cat that is stripy*). In the third block they were asked to try to recall all features (e.g., *a type of cat that is stripy and bluish-grey*). This was done to encourage gradual building up of novel word knowledge, which may result in better learning.

The sentence plausibility judgement task was identical to Experiment 4, with three presentations of each novel word in a compatible sentence, and one in an incompatible sentence. When a meaningless novel word was seen, the participant was instructed to press a key labelled “No meaning”. Each meaningless novel word was assigned four sentences that all suggested a different meaning for the final word. Thus no coherent meaning could be inferred under these conditions. The order of all trials was randomised. E-prime was used to run all tasks using the same equipment as Experiment 4.

Testing. The test session was completed on day 2 after the second training session. Participants were offered a chance to take a rest break before starting the tests. Test tasks consisted of shadowing, reading aloud, cued recall, and meaning recall. The parameters of the shadowing tasks were slightly changed from Experiment 4, in order to make it as similar as possible to the reading aloud task. The test session started with either shadowing or reading aloud. A shadowing trial began with the presentation of a fixation cross for 500 ms. At the offset of the cross, a spoken word was presented through headphones. This started the timing. The trial finished 1500 ms after a vocal response was detected or after 2000 ms from the onset of the spoken word. In the reading aloud task the fixation cross was replaced by a word written in lower case letters, in black Times New Roman font on a white background. The task was to read the word aloud as soon and as accurately as possible. Same timeout was used as in shadowing. The order of the shadowing and reading aloud tasks was counterbalanced. Before starting the shadowing or reading

aloud task, participants completed five practice trials of each task with stimuli not used in the experimental trials. The experimenter remained in the room during practice to ensure that participants understood the tasks and were speaking loud enough for the response to be detected. The shadowing and reading aloud tasks included all 68 trained novel words, 34 real words (same words as in Experiment 4), and 34 untrained novel words. The order of the trials was randomised by the software.

The third task was cued recall. A cue (e.g., *f _ _ k _ _ n*) was presented on screen, and participants were asked to type in the complete word. No feedback was provided, and no time limit was used. This task included all trained novel words, and the untrained novel words encountered in the shadowing and reading aloud tasks.

The last task of the session was meaning recall. All trained novel words were presented on screen in random order, and the task was to type in the complete meaning. If it was a meaningless novel word, participants typed in "No meaning". No time limit was imposed, and no feedback given. DMDX was used for stimulus presentation and response collection in all test tasks. The same computer equipment was used as in Experiment 4.

Participants

Thirty native English speaking students drawn from the University of York and York St. John University populations participated in the experiment (12 male, one left-handed, mean age = 20.6, range = 18-29). No participants reported language disorders. Participants were paid or received course credit, and the most accurate and fastest 50% of the participants were entered into a prize draw for a £10 gift certificate.

4.3.2 Results

Training data

The degree to which participants learned the meanings of the meaningful novel words during training was examined first by looking at meaning recall data. Figure 23 (left y-axis) shows accuracy levels in recall of the objects to which the novel words refer to across the three blocks of this training task. A mixed-effects logistic regression model with subjects and items as random factors, and time of

training (consolidated words on day 1, and unconsolidated words on day 2) and training block (three blocks) as fixed factors was used. Subject-specific slopes for the time of training were added. Contrasts in looking at the interaction between time of training and block showed that the effect of time of training was significantly larger in block 1 compared to block 3 ($b = -0.738$, $z = -3.05$, $p = .002$), and in block 2 compared to block 3 ($b = -0.681$, $z = -2.96$, $p = .003$). The effect of time of training was significant in the first and second blocks ($b = -0.648$, $z = -3.53$, $p < .001$, $b = -0.594$, $z = -3.35$, $p < .001$), but no significant difference between day 1 and day 2 training was found in the third block. Accuracy on day 1 increased with training: accuracy increased from block 1 to block 2 ($b = 1.519$, $z = 9.94$, $p < .001$), from block 1 to block 3 ($b = 2.416$, $z = 14.40$, $p < .001$), and from block 2 to block 3 ($b = 0.900$, $z = 5.54$, $p < .001$). The same was true of day 2 training, where accuracy increased from block 1 to block 2 ($b = 1.563$, $z = 9.86$, $p < .001$), from block 1 to block 3 ($b = 3.147$, $z = 17.44$, $p < .001$), and from block 2 to block 3 ($b = 1.584$, $z = 9.64$, $p < .001$).

The right y-axis of Figure 23 shows the mean number of features participants could recall for the novel word meanings. These data were analysed using ordinal

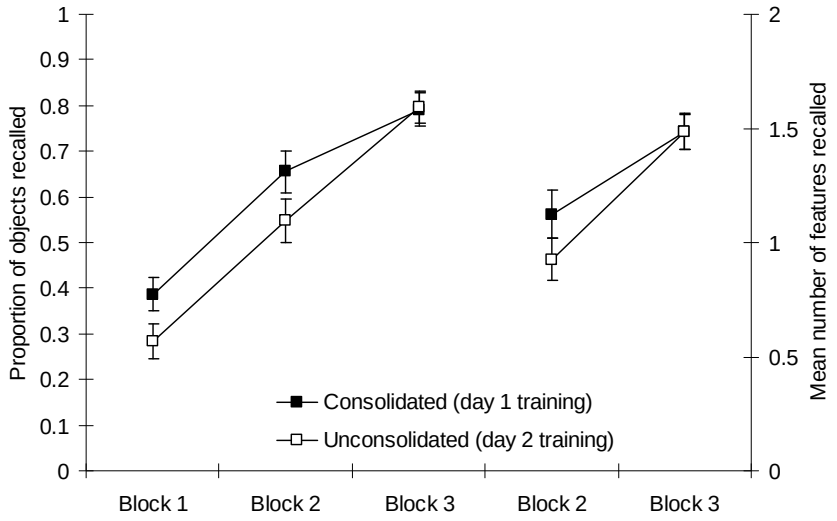


Figure 23. Accuracy rates in the meaning recall training task. Error bars represent standard error of the means.

logistic regression (as in Experiment 1), with time of training (consolidated words on day 1, and unconsolidated words on day 2) and training block (blocks 2 and 3) as predictors (recall that participants were not required to recall features until the

second block). Time of training entered into an interaction with block, showing that the effect of time of training was significantly larger in block 2 than in block 3 ($b = 0.383$, $z = 2.14$, $p = .03^{\dagger}$). The effect of time of training was significant in block 2 ($b = -0.405$, $z = -3.40$, $p < .001$), but not in block 3. Recall accuracy increased from block 2 to block 3 on both day 1 ($b = 0.836$, $z = 6.55$, $p < .001$) and day2 ($b = 1.219$, $z = 9.67$, $p < .001$). Together these object and feature recall accuracy data show that by the end of training, both novel word sets had been learned equally well.

The cued recall data were analysed next to assess the degree to which forms of the meaningless novel words were acquired during training. Figure 24 shows accuracy rates in this training task for both meaningless and meaningful novel words. The data were analysed using a mixed-effects logistic regression model with subjects and items as random factors, and time of training (consolidated words on day 1, and unconsolidated words on day 2), meaningfulness (meaningless vs. meaningful), and training block (three blocks) as fixed factors, which also benefitted from subject-specific slopes for block. No three-way or two-way interactions reached significance. The simplified model showed that, averaged across the other variables, fewer meaningless novel words were recalled correctly than meaningful words ($b = -0.505$, $z = -8.37$, $p < .001$). There was also an effect of time of training, with more words recalled correctly on day 1 ($b = -0.180$, $z = -2.99$, $p = .003$). Also, averaged across the other variables, performance differed between blocks, with

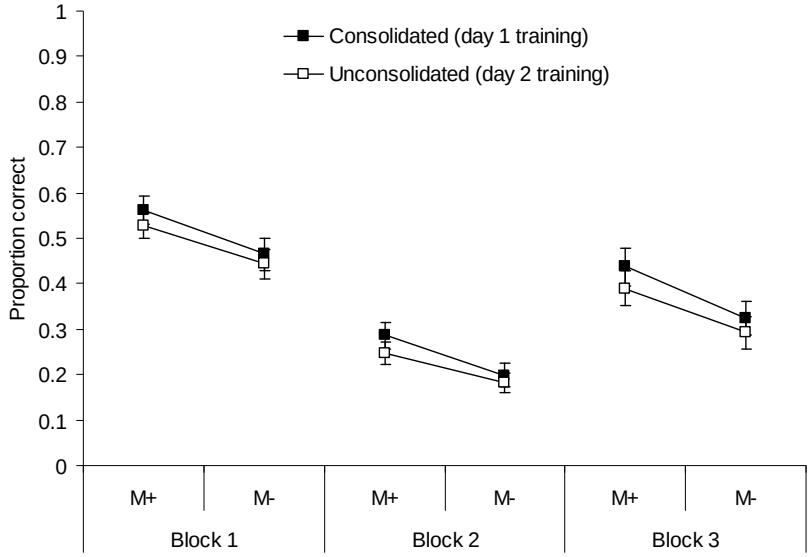


Figure 24. Accuracy rates in the cued recall training task. M+ refers to meaningful novel words, M- to meaningless novel words. Error bars represent standard error of the means.

better recall in block 1 than block 2 (b = -1.473, z = -17.60, p < .001) or block 3 (b = -0.722, z = -6.43, p < .001). This is because the cue used in the first block provided all the letters of the word apart from one, making it a very easy condition. In blocks 2 and 3 half of the letters were missing. There was still significant improvement from block 2 to block 3 (b = 0.751, z = 8.04, p < .001).

Test data

Shadowing. Same data trimming procedure was applied on the shadowing data as was done in Experiment 4. Voice key trigger accuracy was again checked with CheckVocal and corrected when necessary. The shadowing latencies and response accuracy data are shown in Figure 25. A mixed-effects linear model with subjects and items as random variables, and novel word condition (consolidated-meaningful, consolidated-meaningless, unconsolidated-meaningful, unconsolidated-meaningless, untrained, real words) as the fixed variable was fitted. Subject-specific slopes for trial order increased the goodness of fit. Contrasts involving the novel word conditions were examined first. There was no evidence of a consolidation effect: latencies to both meaningful and meaningless novel words were unaffected by the consolidation condition. There was also no evidence of a meaningfulness benefit: in both consolidated and unconsolidated novel words meaningless and meaningful items were shadowed equally fast. The two control conditions differed significantly from the novel word conditions. Untrained novel words were shadowed slower than any other condition (contrast with real words: b = -0.128, t = -8.98, p < .001, consolidated-meaningful: b = -0.058, t = -8.47, p < .001, consolidated-meaningless: b = -0.048, t = -7.01, p < .001, unconsolidated-meaningful: b = -0.048, t = -7.01, p < .001, unconsolidated-meaningless: b = -0.045, t = -6.58, p < .001). Real words on the other hand were shadowed faster than any other condition (contrast with consolidated-meaningful: b = 0.070, t = 4.75, p < .001, consolidated-meaningless: b = 0.080, t = 5.41, p < .001, unconsolidated-meaningful: b = 0.080, t = 5.42, p < .001, unconsolidated-meaningless: b = 0.083, t = 5.59, p < .001).

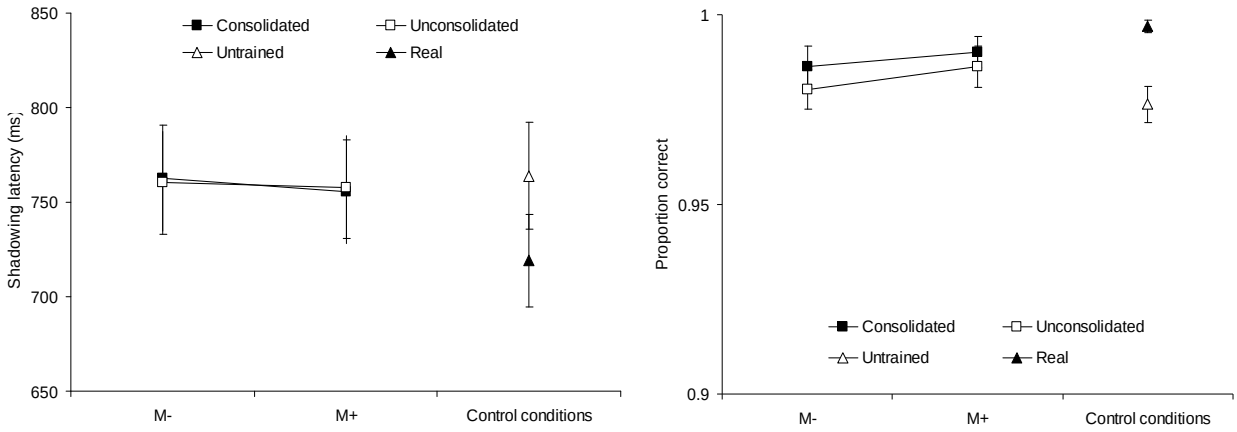


Figure 25. Shadowing latencies and accuracy rates. M+ refers to meaningful novel words, M- to meaningless novel words. Error bars represent standard error of the means.

Accuracy data in the shadowing task are presented in the right panel of Figure 25. The data were analysed using a mixed-effects logistic regression model with subjects and items as random factors, and word condition (consolidated-meaningful, consolidated-meaningless, unconsolidated-meaningful, unconsolidated-meaningless, untrained, real word) as the fixed variable. No effect of consolidation was found, either in meaningful or meaningless novel words. No effect of meaningfulness was found either in the two consolidation groups. Untrained novel words were repeated as accurately as all novel word conditions. The real word condition had a significantly higher accuracy rate than untrained novel words ($b = -2.065$, $z = -2.93$, $p = .003$) and meaningless unconsolidated novel words ($b = -1.878$, $z = -2.49$, $p = .01^{\dagger}$). The difference between real words and meaningful unconsolidated novel words was only marginally significant ($b = -1.497$, $z = -1.90$, $p = .06^{\dagger}$), as was the difference between real words and consolidated meaningless novel words ($b = -1.512$, $z = -1.92$, $p = .05^{\dagger}$).

Reading aloud. The same trimming procedure was carried out on the data as in the shadowing task. Figure 26 (left panel) shows reading latencies in each condition. CheckVocal was used to make sure the latencies were accurately recorded. The analysis was carried out using a mixed-effects linear model with subjects and items as random variables, and word condition (consolidated-meaningful, consolidated-meaningless, unconsolidated-meaningful, unconsolidated-meaningless, untrained, real word) as the fixed variable, and subject-specific slopes

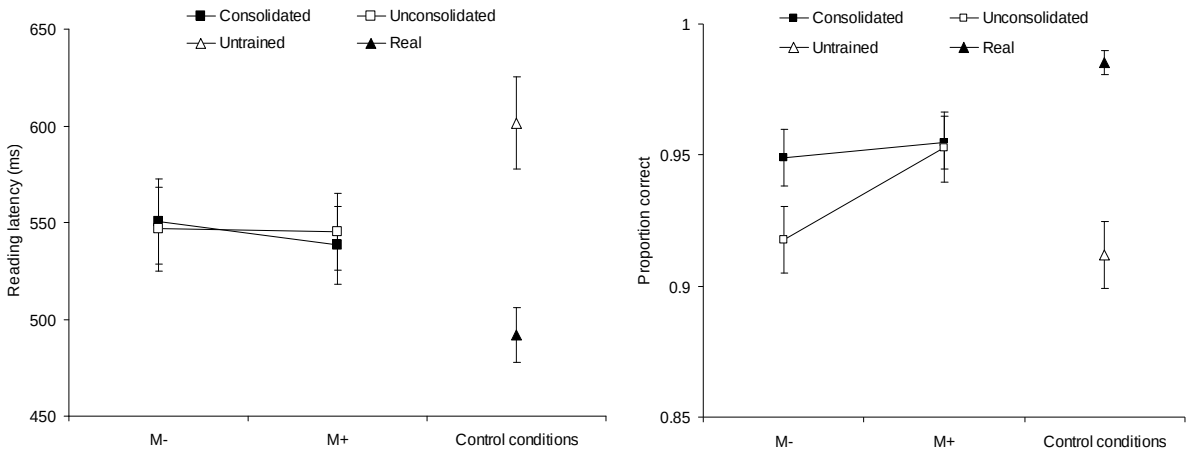


Figure 26. Reading latencies and accuracy rates. M+ refers to meaningful novel words, M- to meaningless novel words. Error bars represent standard error of the means.

for trial order. The effect of time of training did not reach significance for either meaningful or meaningless novel words. There was a marginally significant benefit for meaningful novel words over meaningless novel words in the consolidated condition (b = 0.017, t = 1.80, p = .067†). No such effect was seen in the unconsolidated condition. Reading times to untrained novel words were slower than in any other condition (contrast with real words: b = -0.200, t = -9.94, p < .001, consolidated-meaningful: b = -0.109, t = -13.05, p < .001, consolidated-meaningless: b = -0.092, t = -10.94, p < .001, unconsolidated-meaningful: b = -0.096, t = -11.44, p < .001, unconsolidated-meaningless: b = -0.094, t = -11.12, p < .001). Reading times to real words were significantly faster than reading times in any other condition (contrast with consolidated-meaningful: b = 0.092, t = 4.44, p < .001, consolidated-meaningless: b = 0.110, t = 5.27, p < .001, unconsolidated- meaningful: b = 0.105, t = 5.08, p < .001, unconsolidated-meaningless: b = 0.107, t = 5.14, p < .001).

Accuracy data in the reading aloud task are presented in the right panel of Figure 26. The data were analysed using a mixed-effects logistic regression model with subjects and items as random factors, and word condition (consolidated-meaningful, consolidated-meaningless, unconsolidated-meaningful, unconsolidated-meaningless, untrained, real word) as the fixed variable. There was no effect of time of training in the meaningful condition, but in the meaningless condition consolidated novel words had a significantly higher accuracy rate than unconsolidated words (b = -0.636, z = -2.28, p = .02†). There was no effect of

meaningfulness in the consolidated condition, but in the unconsolidated condition meaningful novel words had a significantly higher accuracy rate (b = -0.747, z = -2.61, p = .01[†]). Untrained novel words had a significantly lower accuracy rate than real words (b = 2.180, z = 5.05, p < .001) or either of the consolidated novel word conditions (meaningful: b = 0.862, z = 3.29, p = .001, meaningless: b = 0.684, z = 2.77, p = .006[†]), or the meaningful unconsolidated condition (b = 0.802, z = 3.12, p = .002). The difference between untrained words and unconsolidated meaningless novel words was non-significant. Real words had a higher accuracy rate than any other condition (consolidated-meaningful: b = -1.320, z = -2.78, p < .001, consolidated-meaningless: b = -1.495, z = -3.20, p < .001, unconsolidated-meaningful: b = -1.380, z = -2.92, p < .001, unconsolidated-meaningless: b = -2.131, z = -4.74, p < .001).

Cued recall. Accuracy rates in the cued recall task are shown in Figure 27. Meaningfulness (meaningful vs. meaningless) and time of testing (delayed = consolidated, immediate = unconsolidated) were entered as fixed factors in a mixed-effects logistic regression model with subjects and items as random variables. Time of testing and meaningfulness did not enter into an interaction. Averaged across the meaningfulness conditions, significantly more consolidated novel words were recalled accurately compared to unconsolidated words (b = -0.571, z = -5.34, p < .001). Averaged across the time of testing conditions, meaningful novel words had significantly higher accuracy rates than meaningless novel words (b = 0.535,

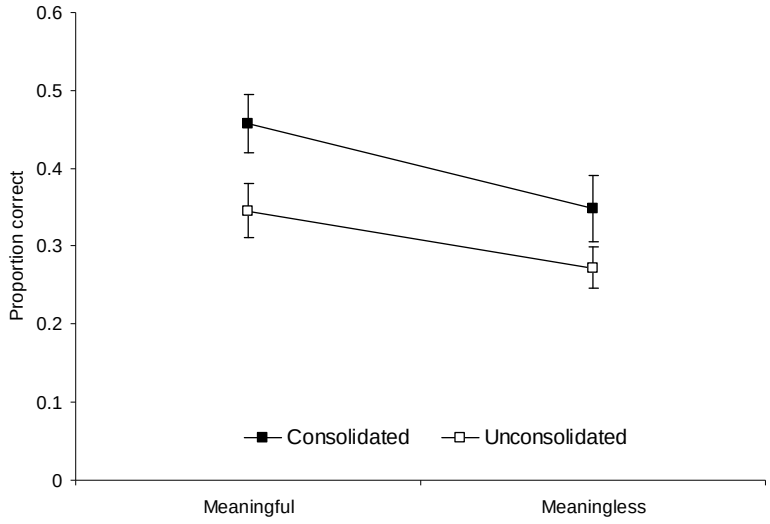


Figure 27. Accuracy rates in the cued recall test task. Error bars represent standard error of the means.

$z = 5.03$, $p < .001$). The effect of time of training was significant also when evaluated at both meaningfulness levels separately (meaningful: $b = -0.660$, $z = -4.49$, $p < .001$, meaningless: $b = -0.468$, $z = -3.03$, $p = .002$). The effect of meaningfulness was also significant when evaluated at both time of training conditions (consolidated: $b = 0.625$, $z = 4.26$, $p < .001$, unconsolidated: $b = 0.438$, $z = 2.84$, $p = .004$).

Meaning recall. Accuracy rates in recalling the meaning of the novel words are displayed in Figure 28, both for recalling the object to which the novel word refers to (left y-axis), and the number of features recalled (right y-axis). Accuracy in recalling objects was analysed with a mixed-effects logistic regression, with subjects and items as random variables, and time of testing as the fixed variable. Subject-specific slopes for time of testing improved goodness of fit. Unconsolidated novel word meanings were recalled significantly more accurately than consolidated words ($b = 1.143$, $z = 4.78$, $p < .001$). The same pattern was seen in the number of features recalled, when analysed with ordinal logistic regression, with more features recalled in the unconsolidated condition ($b = 0.672$, $z = 4.94$, $p < .001$).

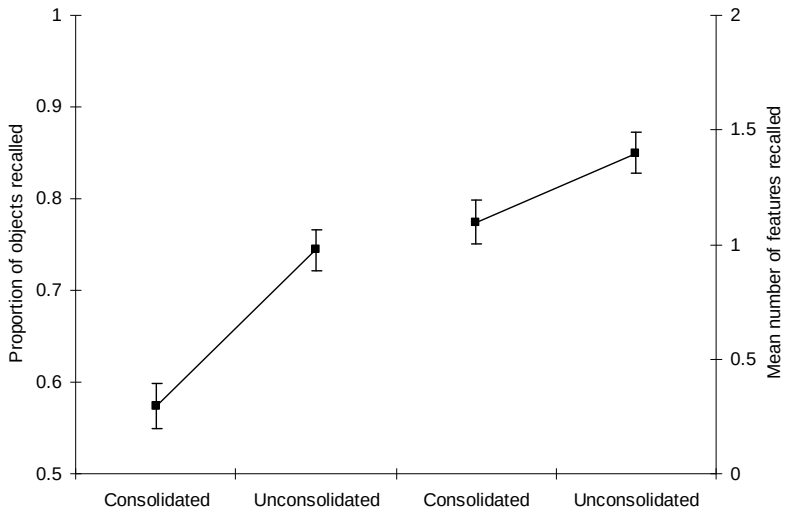


Figure 28. Accuracy rates in the meaning recall test task. Error bars represent standard error of the means.

4.3.3 Discussion

The shadowing task in the current experiment failed to replicate the consolidation effect seen in Experiment 4. Now participants were equally fast and accurate in shadowing novel words they had learned just before testing, and the day

before testing. No effect of meaningfulness was found either, participants shadowed meaningful and meaningless novel words equally fast and accurately. The failure to replicate the consolidation effect was puzzling, considering that the task and the stimuli were identical to Experiment 4, apart from the elaborated meanings. Two other differences may have influenced the result. Firstly, in Experiment 5 the shadowing task was carried out immediately after the day 2 training session (although preceded by the reading task in half of the participants). It is possible that in Experiment 4 the additional exposure provided by the semantic decision, sentence plausibility, and meaning recall tasks acted to raise the activation of the novel word representations and created better circumstances for perhaps weak consolidation effects to reach an observable level. Learning the unconsolidated set of novel words may have interfered with the recall of consolidated novel words, and such short-term interference may have been overcome by a small number of presentations of the novel words before the shadowing task took place. In light of these considerations, a second replication of the shadowing task will be presented in the next chapter, where it is again carried out as the last task of the test session.

The reading aloud task also failed to show consolidation effects in reading latencies. Consolidated and unconsolidated novel words were read equally fast, as were meaningful and meaningless novel words. The reading accuracy rates however did show an interaction between consolidation and meaningfulness. No consolidation effect was seen in meaningful novel words, but in the meaningless condition consolidated novel words were read significantly more accurately than unconsolidated words. This rules out a semantic locus for the consolidation effects in this task: if the effect was caused by consolidation of word meanings, the effect would have been observed in the meaningful condition rather than the meaningless. It may be the case that unconsolidated novel words, being recently acquired and still relatively weak, benefit from the additional boost semantic activation provides. As they become stronger during the course of offline consolidation, the semantic boost becomes less prominent. This view is supported by the observation that the consolidated novel words were unaffected by the meaningfulness manipulation, while the unconsolidated words significantly benefitted from meaning. However, these data should be interpreted with some degree of caution as the effect was seen only in the accuracy rates, but not in reading times. Under the priming hypothesis discussed in connection with Experiment 4, the lack of a reading time consolidation

effect might be taken as evidence that only cross-modal priming benefits from consolidation. However, the failure to replicate the shadowing consolidation effect (cross-modal) makes it difficult to interpret this finding.

The cued recall task was successful in showing both an effect of consolidation and meaningfulness, although no interaction between the two. In the recall task consolidated novel words were recalled significantly better than unconsolidated words, and this was the case for both meaningful and meaningless novel words. These data support the idea that knowledge of novel word forms benefits from offline consolidation, and agree with the shadowing data from Experiment 4, as well as the consolidation effects of novel word forms shown by the lexical competition studies of Gaskell and colleagues (e.g., Dumay & Gaskell, 2007). The effect here does not seem to have a semantic locus, as the consolidation benefit was seen in both meaningless and meaningful novel words. It is somewhat surprising that the similar cued recall task in Experiment 1 did not show a consolidation effect. The overall accuracy rates in the two experiments are similar (when compared with the non-neighbour condition of Experiment 1, which is most comparable to the current experiment), suggesting that the difficulty rates in the two tasks were also similar. One major difference between the two experiments was the design. In Experiment 1 two different groups of participants were tested, one immediately after training, and the other one day later. In Experiment 5 on the other hand the same participants were tested in both consolidation conditions. It may be that the more powerful design of the current experiment allowed the effect to emerge. An alternative explanation is offered by an examination of the training data. In the cued recall training task performance on day 1 was slightly but statistically significantly better than in the day 2 training session. Thus it is possible that participants were more attentive or motivated in the first training session than in the second, and learned the first set of novel words (consolidated) better. It should be noted though that the difference between the two word sets in testing was about three times larger than during training (difference during training averaged across blocks and meaningfulness was 3.1%, in the testing session it was 9.4%, averaged across meaningfulness conditions), suggesting that the difference in training performance may not be the only source of the effect.

In the cued recall task meaningful novel words were recalled better than meaningless novel words. This was true in both consolidation conditions. This

indicates that novel word forms are easier to learn when they carry meaning. A similar conclusion was reached in Experiment 1, where novel words whose form and meaning were semantically related resulted in better cued recall performance compared to words whose form and meaning were unrelated. However, it is important to acknowledge that the meaningfulness effect in the current experiment may have been partially affected by training performance, in that participants may have spent more time learning the meaningful novel words as these items required learning of both form and memory. This extra effort may have caused better learning of forms.

Finally, the meaning recall data replicate findings from Experiment 4. Meanings of the recently learned unconsolidated novel words were recalled significantly better than meanings of consolidated words. This was seen in both recall of objects and features. It appears that in this task the meanings of the novel words learned on the previous day are subject to forgetting or interference from the meanings of the second set of novel words.

4.4 Chapter Summary and General Discussion

The aims of the two experiments described in this chapter were to evaluate offline consolidation effects in access to novel word meanings and novel word forms. Experiment 4 focused on meaning, using a semantic decision task as the main test of meaning access. When novel words acted as primes, participants were faster to make a semantic decision about the relatedness of the prime and target when the two were semantically associated than when they were unassociated. This replicated earlier reports of semantic decision performance using novel word primes (e.g., Perfetti et al., 2005). Importantly though, the experiment reported here was the first time consolidated and unconsolidated novel words have been compared in this task, and showed a similar relatedness effect in both conditions. The similar performance in the two time of testing conditions might be interpreted as evidence against a role for consolidation, however such a conclusion would be premature. Firstly, towards the end of this task decision latencies to trials with consolidated novel words had become faster than latencies to trials with unconsolidated novel words (although this was seen in the related condition only). Secondly, the sentence plausibility judgement task, which is similar to the semantic decision task in that both require a

decision to be made about the semantic fit of the novel word in a provided context, showed a significantly larger compatibility effect (difference between trials where the novel word fits and trials where it does not fit with the sentence context) with consolidated than unconsolidated novel words (although see discussion below for an alternative view of this task). Together these tasks suggest that meanings of consolidated novel words may be accessed faster than meanings of unconsolidated novel words.

In contrast to these tasks showing a consolidation benefit in speed of access to meaning, both Experiment 4 and 5 showed that in explicit meaning recall there was a benefit for unconsolidated novel words. The dissociation between these two types of task is interesting. Meaning recall requires an explicit act of retrieval of the novel word meaning, in as much detail as possible. This task does not appear to benefit from consolidation. The semantic decision and sentence plausibility tasks on the other hand require a speeded decision about the meaning of the novel word in relation to an existing word or a semantic context. It is these two tasks which show some evidence of consolidation. This dissociation makes sense in the CLS view, where one of the most important functions offline consolidation has is to connect and interleave new information with existing information. Since the semantic decision and sentence tasks are the only tasks which require participants to relate known words with new words, these tasks are the most likely ones to be sensitive to consolidation processes. In other words, these semantic tasks provide preliminary evidence that novel word meanings are gradually integrated with existing semantic knowledge.

I next turn to consider the tasks measuring novel word form recall. Experiment 4 included a shadowing task which showed a clear advantage for consolidated novel words in shadowing latencies and accuracy rates. While this finding was not replicated in Experiment 5, the latter experiment did provide further evidence for consolidation of novel word form in other tasks. Accuracy rates in the reading aloud task showed a consolidation advantage for meaningless novel words (although the effect was not seen in reading latencies). The cued recall data supported this pattern, by showing better recall of consolidated novel word forms, both when they were meaningful and meaningless. These demonstrations of consolidation of novel word forms in the visual modality join the growing literature on auditory novel word lexical competition research, which has shown a crucial role

for consolidation in integrating novel words in the mental lexicon (e.g., Dumay & Gaskell, 2007). It is important to note however that the cued recall consolidation data were in conflict with the null finding in Experiment 1. The cued recall task will be revisited in Chapter 6, using a more sophisticated design disentangling sleep and wake-related consolidation.

Experiment 5 also attempted to clarify the role of meaning in novel word learning. There was some evidence for a facilitatory role for meaning. Meaningful novel words were recalled more accurately in the cued recall task, although this may have been caused by an attentional effect during training, with more attention (or effort) allocated to novel words for which a meaning had to be acquired. In the reading accuracy data meaning had a significant benefit, but only in the unconsolidated words. This latter finding may help to explain some of the discrepancies in the reading literature with regard to reading newly learned words. As reviewed in Chapter 1, some studies have failed to find a meaning advantage (e.g., Nation et al., 2007; McKague et al., 2001). However, these studies included multiple training sessions over several days, followed by a later test session. If meaning effects are most prominent in unconsolidated novel words, then any studies attempting to find a meaning benefit in consolidated stimuli would struggle to find one. In fact, McKay et al. (2008) who tested their participants immediately after training did find a meaning advantage in some conditions (see Chapter 1 for details). Although the current experiment does not resolve this issue, it does strongly imply that consolidation is an important factor to consider in novel word reading experiments.

The evidence for consolidation of word forms recommends a re-evaluation of the data from Experiment 4. Although both the semantic decision task and the sentence plausibility tasks were intended to measure access to word meaning knowledge, both tasks are likely to be heavily influenced by form knowledge too. This is particularly the case in the sentence task. In this task participants were allowed to view the sentence without the concluding novel word for as long as they wanted. This was done to allow for variability in reading speed. However, it is possible that participants had the time to generate a guess about the final word. The sentences were designed to be highly constraining, so guessing the identity of the missing novel word should not have been difficult. This means that when the novel word was revealed, the participants' task would simply have been to confirm that the

appearing novel word was indeed the same word they were expecting. In these trials task performance would then have been more based on recognition of form than meaning. It is less plausible to suggest that the same strategy would have worked in the semantic decision task, but the evidence for consolidation was weaker in that task anyway. To assess this alternative explanation of the sentence plausibility task, Experiment 6 used a modified version of the task where a rapid serial visual presentation (RSVP) method was used to reduce the opportunity to use a guessing strategy.

The semantic decision task may not be the optimal task to evaluate speeded access to novel word meanings either, as it requires an explicit decision to be made about the identity of the novel word. As discussed earlier, offline consolidation is likely to have the strongest effect in tasks which measure the degree to which the novel word meaning has been integrated with existing semantic structures. If this is the case, then the most sensitive tasks for semantic consolidation effects would be traditional semantic priming paradigms, where no explicit response is made to the novel word itself, but where the influence of the novel word is measured in access to a semantically related familiar word. These tasks also provide a purer measure of semantic activation, with less potential confound from form based processing. The two experiments reported in the next chapter (Experiments 6 and 7) will look at semantic priming in a commonly used primed lexical decision task, where the prime is a novel meaningful word. These tasks should clarify the conclusions from the semantic decision task. In sum, although the experiments reported in this chapter provide preliminary evidence for a possible consolidation benefit in learning novel word meanings, more data are needed from tasks which measure online activation of semantics. Experiment 5 also showed that offline consolidation plays a role in learning of novel word forms. This process will be examined in more detail in Experiment 8.

Chapter 5: Semantic priming using newly learned meaningful words

5.1 Introduction

Experiment 4 suggested that when asked to make speeded decisions about the meaning of a novel word, people were faster to respond to novel words that had had a chance to consolidate over a 24 hour period, compared to novel words that had been learned briefly before testing. This implies faster access to the meanings of consolidated novel words, which interestingly was not reflected in increased recall accuracy in an untimed meaning recall task. However, as discussed in the previous chapter, this effect was only seen in the last block of the semantic decision task. Also, the semantic decision task measures speed of explicit retrieval of the novel word meaning by requiring an explicit decision to be made about the identity of the novel word and its relationship to the prime. The experiments reported in this chapter made use of semantic priming as a task not requiring an explicit response to the novel word, and hence tapping into more automatic semantic activation. This type of access may be more sensitive to potential consolidation effects, and is more likely to reflect normal online semantic processing by not requiring participants to explicitly retrieve the novel word meanings.

Semantic priming, as discussed in earlier chapters, refers to the finding that when participants are asked to recognise a word (e.g., *doctor*), typically in a lexical decision or a naming task, they respond faster when the target word is preceded by a semantically related or associated word (e.g., *nurse*), compared to an unrelated word (e.g., *tiger*) (see Neely, 1991, and McNamara, 2005, for comprehensive reviews). The most common explanation for priming is based on spreading semantic activation. One often cited model of semantic priming that relies on spreading activation was proposed by Collins and Loftus (1975). According to this view knowledge is represented as a semantic network consisting of interconnected nodes. Each node represents a semantic concept, and is surrounded and connected to related concepts. When a prime word is encountered, it activates its corresponding concepts. This activation then spreads to the related, connected nodes. If a related target word is presented briefly afterwards (as activation decays over time), the residual activation from the prime will facilitate the activation and recognition of the target.

Many other models rely on similar processes of spreading activation. In distributed network models (e.g., Plaut & Booth, 2000) a concept is not represented as a node in a network, but rather as a pattern of activation over a number of units. Related concepts share semantic features, and hence also have overlapping representations. As a prime is presented, it activates its corresponding pattern of units. When a related target is subsequently presented, it can be activated faster because some of its units were already active when the prime was processed.

Multistage activation models share the operating principles of the above models, but add different stages of lexical-semantic processing. The interactive-activation model of Stolz and Besner (1996) applies to visual word recognition, and consists of letter, lexical, and semantic levels. The presentation of a prime word activates the relevant letters at the letter level. Activation feeds forward to the lexical level, where lexical representations that match the incoming letter information are activated. Activation of a lexical representation further feeds to the semantic level. Importantly though, activation at the semantic level is not restricted to the one semantic representation that best matches the incoming information, but also applies to representations that are semantically related. The activation of all of these semantic representations feeds back to the lower levels, ensuring that words related to the prime will also be activated at all lower levels, although less strongly. The related target can then be processed more efficiently at each level than an unrelated target.

The spread of semantic activation is often considered to be an automatic process, in that it operates outside of consciousness and without effort. However, semantic priming is also affected by strategic processes. Some models of semantic priming explicitly incorporate such factors. The three-process theory proposed by Neely and Keefe (1989) incorporates spreading activation that occurs in an automatic fashion, but also includes an expectancy process whereby participants under the right circumstances can explicitly generate hypotheses about the identity of the target based on the prime. The third process in this model is a semantic matching process where participants search for a relation between the prime and target in order to facilitate the word/nonword decision made in response to the target (this process only applies in primed lexical decision). If a relation is found, the response must necessarily be “word”, whereas the absence of a relation may prime a “nonword” response. The degree to which this third process is helpful depends on the proportion

of trials where the prime and target are related, known as relatedness proportion (RP), with a high RP encouraging more strategic processing. The proportion of nonword trials (nonword ratio, or NR) is also relevant. Under conditions of a high NR the absence of a semantic relation is likely to signal a nonword. Under a low NR on the other hand the absence of a relation is more likely to signal a word. By manipulating the RP and NR the experimenter can make participants' response strategies more or less reliant on automatic or strategic processing. A third variable that is relevant here is the SOA between prime and target. A short SOA will not allow time for expectancy processes, again biasing the task towards more automatic processing.

The experiments reported in this chapter were carried out to see if novel words enable semantic priming, showing that the novel word has been integrated in the semantic system. The second aim was to see if offline consolidation of novel word meanings is necessary for priming to emerge, and if it is equally important for strategic processing of meaning and automatic processing of meaning. There are no commonly agreed rules on how to elicit automatic as opposed to strategic priming, and it is likely that all tasks involve both processes to varying degrees. Hence the aim here was to design priming experiments that were likely to fall either towards to strategic or automatic end of the continuum. Experiment 6 was intended to tap into more strategic processing by using a visible prime and a long SOA, and Experiment 7 was intended to tap into more automatic processing by using a masked prime and a short SOA.

The two experiments evaluated consolidation that takes place within 24 hours of learning as well as consolidation that may take place over a longer time course of several days and nights, by testing priming immediately after training, one day after training, and one week after training. Because in Experiment 6 the novel word primes were visible and participants were aware of their potential semantic relationship with the targets, an experimental design was needed which does not involve repeated exposure to the words across the test times. This was because such repeated exposure might act as a further training opportunity. Hence in Experiment 6 one group of participants was tested on words learned on the day of testing and on another set of words learned one day before testing. A different group of participants was tested on words learned on the day of testing and on another set of words learned one week before testing. The ideal design would be to compare short

consolidation and long consolidation in the same participants as this optimises statistical power. While this was not possible in Experiment 6 for the reasons stated above, it was possible in Experiment 7 which used masked primes. In these circumstances participants are not explicitly exposed to the novel word primes, and there is less risk of additional explicit training taking place. Hence in Experiment 7 the same group of participants did the priming task three times, once immediately after training, again one day after training, and once more seven days after training.

5.2 Experiment 6

Experiment 6 evaluated the ability of newly learned words to prime familiar associated words. Note that the relationship between a prime and a target can be defined in two different ways: they can either be semantically related, in that they share semantic features (e.g., dog-goat, where both are mammals, have fur etc.), or they can be semantically associated, in that they often occur together in a similar context, and often occur together in free association tasks (e.g., dog-cat). As can be seen in the examples cited above, in most instances priming is a mixture of both relatedness and association, as words that are highly associated often also tend to be semantically related. In a meta-analysis Lucas (2000) concluded that both semantic relatedness and association resulted in priming, but association provided an extra boost in the magnitude of priming. Hutchison (2003) on the other hand argued based on his analysis that there was no evidence for priming in the absence of association. The experiments reported in this chapter used prime-target pairs that are associated, and thus are likely to reflect both semantic relatedness and association.

The present experiment followed the design used in Experiment 4 with the addition of a second consolidation condition where the difference between learning the consolidated and unconsolidated words was one week instead of one day. Figure 29 illustrates the timing of training and testing sessions. All participants were trained on one set of novel words and their meanings on day 1. Half of the participants returned on day 2 to be trained on a second set of novel words (short consolidation opportunity). After the second training session, a test session was initiated, with tests of explicit meaning recall, semantic priming (primed lexical decision), sentence plausibility, and shadowing. The other half of the participants returned instead on day 8 (long consolidation opportunity), and carried out the same training and testing

tasks as the short consolidation group. The purpose of adding the long consolidation opportunity group was to assess the possibility that semantic information benefits from more than one day or night of consolidation.

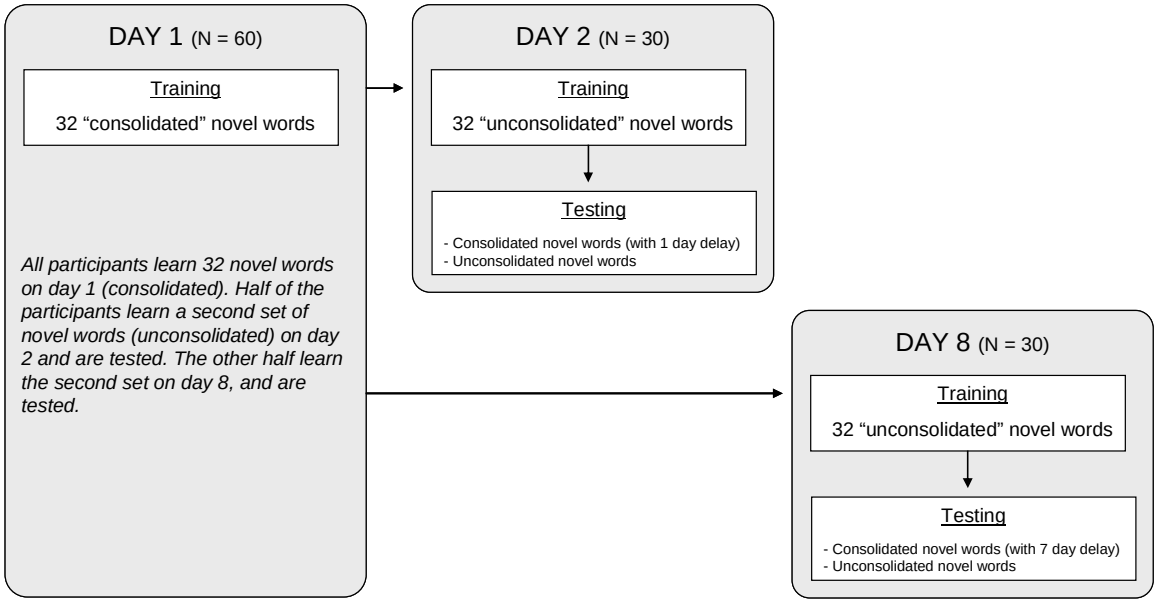


Figure 29. Timing of training and testing sessions in the two consolidation groups (1 day = short consolidation opportunity, 1 week = long consolidation opportunity).

Semantic priming was tested by primed lexical decision, where participants were required to make a word/nonword decision to a target, which was preceded by either an associated or unassociated prime word. This task was administered using both real word primes and novel word primes, in different blocks. The real word prime condition was included to make sure the parameters chosen with regard to RP, NR and SOA resulted in priming. McNamara (2005) has recommended that in order to look at strategic priming one should choose an SOA of over 200 ms, and an RP of over 0.2. Consequently the parameters chosen here included a long SOA of 450 ms, and an RP of 0.5. NR was set at 0.5.

As discussed in the previous chapter, the sentence plausibility task in Experiment 4 may have been affected both by semantic and orthographic processes, due to the self-paced nature of the task. The task was modified for the current experiment by using a rapid serial visual presentation (RSVP) paradigm where the sentence was presented one word at a time at a fixed speed, and participants were asked to respond to the novel word, which always occurred at the end of the

sentence, as quickly as possible. This presentation method should reduce the chances of generating an explicit guess about the identity of the novel word, and should measure speed of semantic access with less influence from purely form-based processes.

Finally, the shadowing task was included in exactly the same format as in Experiment 4, to see if it could be replicated here, after the unsuccessful replication in Experiment 5. If the failure to replicate this effect in Experiment 5 was due to introducing it immediately after the training, it should emerge again in this experiment where it is again carried out at the end of the session. If on the other hand the effect is genuinely unreliable, it should not be seen here either. Also, it is possible that this task too benefits from consolidation over more than one day or night. If this is the case, consolidation effects should emerge only in the long consolidation opportunity group.

5.2.1 Method

Materials

The novel words were the same 102 nonwords used in Experiments 4 and 5. The novel word meanings consisted of the same 68 objects with two features as used in Experiment 5. Care was taken to make sure that no features overlapped with the targets used in the priming test. The real word targets used in primed lexical decision with novel word primes were the 204 targets from Experiment 4 (three associated targets for each novel word meaning). The 34 real word primes and 102 real word targets in the real word priming control condition were also taken from Experiment 4. The properties of all these stimuli were described in Chapter 4.

The primed lexical decision task required generation of 204 nonword targets. These were created by changing one letter in the real word targets to generate legal nonwords (e.g., *strepe* derived from *stripe*). The motivations for this procedure were to generate nonword targets that were carefully matched with the word targets, and to make sure the nonwords were word-like in order to increase task difficulty to make sure participants would be more likely to benefit from the associated primes. Roughly equal numbers of nonwords were created by changing letters in all positions of the words, with consonants always replaced with consonants, and vowels with vowels. The nonwords are presented in Appendices 5 and 6. For the

purposes of the sentence plausibility judgement task in training and testing, the same sentences were used as in Experiment 5.

Design

In Experiment 4, the 102 nonwords were divided into three lists of 34 nonwords each, to be used in the two trained conditions ("consolidated" novel words trained on day 1, "unconsolidated" novel words trained on day 2), and one untrained condition, matched in length. The same lists were used in the current experiment for training on day 1, and on day 2 or day 8. The design of Experiment 4 also divided the 68 meanings into two lists of 34 meanings to be used in the different consolidation conditions (unconsolidated and consolidated), matched in length and frequency. The same lists were used here. The lists were again rotated through all conditions across all participants.

Procedure

Training. Training took place for all participants on day 1, for half of the participants additionally on day 2, and for the other half additionally on day 8. No testing took place on day 1. As in the previous chapter, I shall again refer to words learned on day 1 as "consolidated" novel words (with a further distinction between participant groups with long or short consolidation opportunity), and words learned on the day of testing (day 2 or day 8) as "unconsolidated" novel words, with the same theoretical caveats as before. Training on both days was identical, and consisted of the same tasks used in Experiment 4, namely word-to-meaning matching (five exposures to each word), meaning-to-word matching (five exposures), meaning recall (three exposures), and sentence plausibility judgement (four exposures). The tasks were performed in fixed order. The procedure of these tasks was the same as in Experiment 4.

Testing. The testing session followed from the second training session on day 2 or day 8. Participants were offered a chance to take a rest break before beginning the test, and were given written and verbal instructions. Test tasks consisted of meaning recall, primed lexical decision, sentence plausibility judgement, and shadowing (see Table 3). The first task was meaning recall. In this experiment meaning recall was carried out first in order to allow participants to explicitly access their knowledge of the novel word meanings once before doing the primed lexical

decision task. It was hoped that this would cancel out any possible episodic recency effects that might benefit the unconsolidated words, and result in a purer measure of speed of access to novel word semantics. As before, the meaning recall task required participants to type in the full meaning of a trained novel word presented on screen. No time limit was imposed, and no feedback was provided. The order of trials was randomised by E-prime.

Table 3. Sequence of tasks in the second experimental session (day 2 or day 8) in Experiment 6.

Training session: learn novel words and meanings in 4 training tasks feckton is a type of cat that is bluish-gray and has stripes glain is a type of book that has pictures and is ovesized . . .		
Test 1: meaning recall feckton glain . . .		
Test 2: primed lexical decision (prime – target pairs shown below)		
Block 1	Block 2	Block 3
feckton – kitten *(related)*	feckton – dog *(related)*	feckton – fork *(unrelated)*
feckton – noight *(nonword)*	feckton – suf *(nonword)*	feckton – school *(nonword)*
glain – night *(unrelated)*	glain – sun *(unrelated)*	glain – school *(related)*
glain – ditten *(nonword)*	glain – deg *(nonword)*	glain – firk *(nonword)*
.	.	.
.	.	.
.	.	.
Test 3: sentence plausibility The woman liked to listen to the purring of her feckton *(correct usage)* The businessman kept his suits neatly in his glain *(incorrect usage)* . . .		
Test 4: shadowing (auditory presentation) /feckton/ /glain/ . . .		

Note: On day 1 only the training session was carried out. Different sets of novel words were learned on the two days. In primed lexical decision, real word prime condition is not shown in the table.

A primed lexical decision trial began with the presentation of a fixation cross for 500 ms. This was replaced by the prime in lowercase letters for 200 ms, and then the target also in lowercase letters for 200 ms, with an SOA of 450 ms (ITI 250 ms). Timing started from the onset of the target, and 2000 ms was allowed for a response

to be made. The participant's task was to decide whether the target was a real word or a nonword by pressing a key on a Cedrus button box labelled "Word" or "Nonword". Half of the participants responded to "Word" with their right hand, and the other half with their left hand. Once a response was made, feedback was provided both in terms of response accuracy and response time (by displaying the RT). Note that in Experiment 4 no feedback was given about response accuracy because there participants were assessing the semantic relationship between the novel word and a related or unrelated target, and providing accuracy feedback would have offered a chance for further learning. In the current experiment responses were made about the lexicality of the target and not about the novel word prime, hence feedback could be given to encourage fast and accurate performance. A rest break was offered half way through the trials, together with a summary of accuracy statistics (percentage correct so far). Participants were informed that in many trials the prime and target would be related, to encourage them to attend to the prime as well as the target.

There were two versions of the primed lexical decision task. In the novel word version the prime was always a novel word, and the target was a real word or a nonword. In the real word version the prime was always a real word, and the target was again a real word or a nonword. The order in which the two versions were done was counterbalanced across participants. As in Experiment 4, both versions of the task were divided into three blocks, with each prime occurring twice within each block, once with a word target and once with a nonword target (see Table 3). Since there were an odd number of blocks, it was impossible to balance the primes so that each participant would see each prime an equal number of times with an associated and unassociated word target. However, the stimuli were counterbalanced so that across all participants each prime occurred in each priming condition an equal number of times, and within participants half of the primes appeared twice in the associated condition and once in the unassociated condition, and the other half of the primes appeared once in the associated condition and twice in the unassociated condition. For the targets a typical split-plot design was implemented where each participant saw each word or nonword target only once, but across all participants all targets appeared in both priming conditions an equal number of times. This was achieved by randomly dividing the targets into two lists, and presenting one list with associated primes and the other with unassociated primes.

The order of trials was pseudorandomised using Mix (van Casteren & Davis, 2006). At least 15 trials separated the repetition of any prime. A maximum of three consecutive trials from the same time of training or priming condition were allowed, and a maximum of four consecutive trials from the same lexicality condition were allowed. A new pseudorandomised order was generated for each participant.

Sentence plausibility was the third task in the testing session. A trial started with the presentation of a fixation cross for 500 ms, followed by the presentation of the sentence one word at a time, at the speed of 250 ms per word. Once the last word was presented a response cue was shown ("???") which also started timing. The participant was given a maximum of 2000 ms to judge whether the final novel word fitted the semantic context of the sentence by pressing a key on the Cedrus button box labelled "Yes" or "No". Once a response was made, both accuracy and RT feedback was given. Another key press was required to initiate a new trial. The order of trials was randomised by E-prime. Responses were collected to all 68 trained novel words, with a split-plot design used to elicit one response to each novel word (either semantic match or mismatch) with items rotated through conditions across participants.

The final task of the testing session was shadowing, where participants were asked to repeat an auditorily presented word as soon and as accurately as possible. As before, shadowing responses were produced for all 68 trained novel words, 34 untrained novel words, and 34 real words. The procedure, items, instructions, and equipment used were exactly the same as in Experiment 4.

Participants

Sixty native English speaking participants drawn from the University of York and York St. John University student and staff populations participated in the experiment. Thirty were allocated in the short consolidation opportunity group (11 male, one left-handed, mean age = 21.2, range = 17-42), and 30 in the long consolidation opportunity group (seven male, three left-handed, mean age = 20.6, range = 18-28). No participants reported language disorders, or had participated in the previous experiments. Participants were paid or received course credit. The most accurate and fastest 50% of the participants were entered into a prize draw for a £10 gift certificate.

5.2.2 Results

Training data

Accuracy rates in the meaning recall task during training were analysed to see if participants performed equally well on both days of training. These data are presented in Figure 30 (objects on the left y-axis, features on the right y-axis). The object data were analysed first. A mixed-effects logistic regression model with subjects and items as random factors, and time of training (consolidated words on the first training day, and unconsolidated words on the second training day), training block (three blocks), and length of consolidation opportunity (short or long) as fixed factors was fitted. Subject-specific slopes for the time of training improved goodness of fit. Three-way interaction contrasts showed that in the long consolidation opportunity group the difference between the two time of training conditions changed from block 1 and block 2 to block 3 more than in the short consolidation group ($b = 0.905$, $z = 3.35$, $p < .001$ and $b = 0.609$, $z = 2.28$, $p = .02^{\dagger}$ respectively), suggesting that in the short consolidation group the difference between time of training conditions remained stable across block while in the long consolidation group the difference became attenuated by the third block.

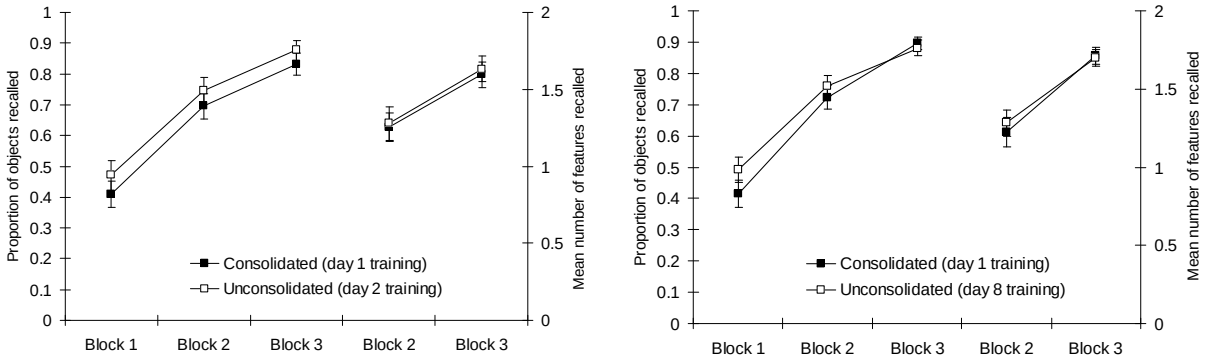


Figure 30. Accuracy rates in the meaning recall training task for both the short (left panel) and the long (right panel) consolidation opportunity groups. Error bars represent standard error of the means.

Since there was a significant difference between the two consolidation length groups, their training data were analysed separately. The object data for the short consolidation opportunity group were analysed using an identical model as the one described above (left panel of Figure 30). Here no significant interaction was found

between time of training and block variables. Averaged over blocks, performance on the second day of training was significantly better (b = 0.496, z = 2.73, p = .006). Accuracy improved from block 1 to block 2 (b = 1.774, z = 21.06, p < .001, from block 1 to block 3 (b = 2.911, z = 29.37, p < .001) and from block 2 to block 3 (b = 1.136, z = 12.06, p < .001). The difference between the two day of training conditions was indeed significant in all three blocks individually (block 1: b = 0.396, z = 1.97, p = .05†, block 2: b = 0.516, z = 2.46, p = .01†, block 3: b = 0.708, z = 3.07, p = .002).

Number of features in the short consolidation group was analysed in the same way as above (using ordinal logistic regression). No interaction was found between the fixed variables. Averaged across the time of training conditions, recall improved significantly from block 2 to block 3 (b = 0.908, z = 13.23, p < .001). There was no overall effect of time of training, and the effect did not reach significance in either block individually. This group then seemed to show some evidence of more effective learning on the second training session, although this effect was restricted to number of objects recalled.

The right panel of Figure 30 shows the training data for the long consolidation opportunity group. Looking at object recall first, an interaction was found between time of training and block, whereby the difference between time of training conditions was smaller in block 3 than either in block 2 (b = 0.439, z = 2.32, p = .02†) or block 1 (b = 0.663, z = 3.48, p < .001). The difference between the two conditions was significant in block 1 (b = 0.411, z = 2.35, p = .02†) but not in block 2 or in block 3. Recall accuracy increased as a function of block in both time of training conditions (first day, block 1 vs. block 2: b = 1.776, z = 15.72, p < .001, block 1 vs. block 3: b = 3.222, z = 22.80, p < .001, block 2 vs. block 3: b = 1.444, z = 10.56, p < .001, second day, block 1 vs. block 2: 1.548, z = 13.96, p < .001, block 1 vs. block 3: b = 2.552, z = 19.54, p < .001, block 2 vs. block 3: b = 1.007, z = 7.64, p < .001). The difference between time of training conditions was not significant when averaged over blocks, but the performance improvement across blocks was significant when averaged over time of training conditions (block 1 vs. block 2: b = 1.656, z = 20.86, p < .001, block 1 vs. block 3: b = 2.882, z = 29.75, p < .001, block 2 vs. block 3: b = 1.225, z = 12.90, p < .001).

Looking at feature recall, no interaction between block and time of training was found. Averaged over block, there was no difference between the time of training conditions. Performance improved from block 2 to block 3 ($b = 1.297$, $z = 18.20$, $p < .001$). The difference between the time of training conditions did not reach significance in either block individually either. Thus in the long consolidation opportunity group there was little evidence of better learning on either training day.

Very high accuracy rates were seen in the sentence plausibility training task with proportion of correct responses at 0.95 on the first and second training days. The difference between the two was non-significant when tested with a mixed-effects logistic regression model with subjects and items as random factors, and time of training (first vs. second testing day) and consolidation length (short and long) as a fixed factors, and there was no interaction with consolidation group.

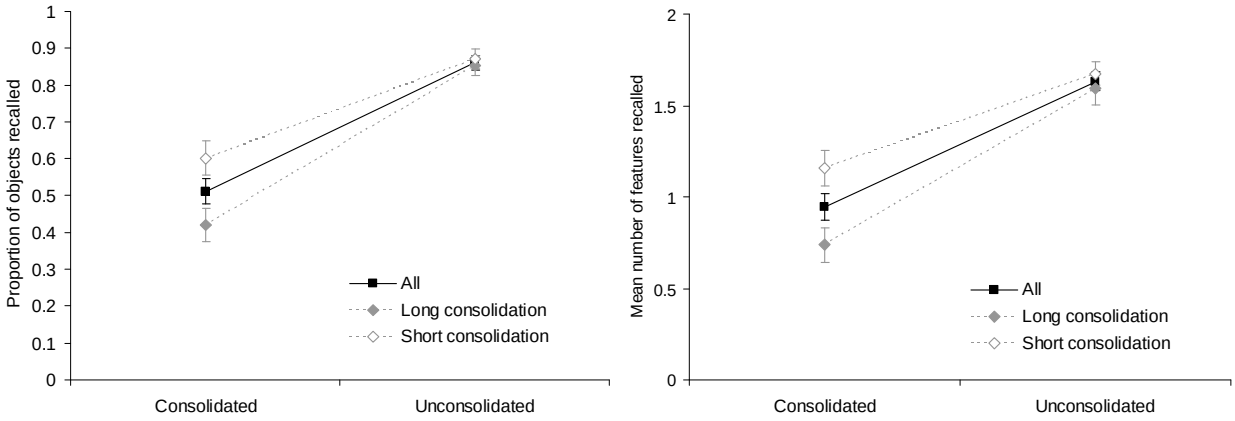


Figure 31. Accuracy rates in the meaning recall test task for both consolidation groups (object recall in the left panel, feature recall in the right panel). Error bars represent standard error of the means.

Test data

Meaning recall. Proportion of novel word meanings recalled in the meaning recall test task are shown in Figure 31, with object recall in the left panel, and feature recall in the right panel. The figure shows the data collapsed across the two consolidation length conditions (solid black line) and separately for the short and long consolidation opportunity conditions (dashed grey lines). The object recall data collapsed across consolidation length conditions were analysed first. A mixed-effects logistic regression model with subjects and items as random variables, and time of testing (immediate = unconsolidated, delayed = consolidated) as the fixed variable was built. The contrast between consolidated and unconsolidated novel word recall

was significant, showing that participants recalled more novel word objects in the words learned immediately before testing ($b = 2.337$, $z = 24.93$, $p < .001$). Figure 31 however suggests that participants in the long consolidation opportunity group recalled fewer consolidated objects. Hence the two consolidation length groups were compared by adding consolidation group as a fixed factor to the model. This revealed a significant interaction between time of testing and consolidation length, whereby the difference between the consolidation length groups was significantly larger in the consolidated than in the unconsolidated condition ($b = 0.716$, $z = 3.84$, $p < .001$). There was no significant difference between the two groups' ability to recall unconsolidated objects, but the short consolidation length group recalled significantly more consolidated objects ($b = 1.007$, $z = 3.07$, $p = .002$) than the long consolidation group. The difference between consolidated and unconsolidated conditions was significant for both groups (long consolidation: $b = 2.665$, $z = 20.39$, $p < .001$, short consolidation: $b = 1.954$, $z = 14.63$, $p < .001$). The conclusion that can be drawn from this analysis is that overall more objects were recalled for words that had been learned on the day of testing, but also that the long consolidation opportunity group seemed to have forgotten more consolidated objects than the short consolidation group.

An ordinal logistic regression model with the same fixed factors showed that, collapsed across the two consolidation length groups, more features were recalled in the words learned immediately before testing ($b = 1.580$, $z = 22.74$, $p < .001$). Again, the two consolidation length groups were compared by adding consolidation group as a fixed factor to the model. A significant interaction was again found between consolidation condition and consolidation length ($b = 0.616$, $z = 4.40$, $p < .001$). Here the short consolidation group recalled significantly more features of both consolidated ($b = 0.862$, $z = 9.74$, $p < .001$) and unconsolidated novel words ($b = 0.246$, $z = 2.27$, $p = .02^{\dagger}$). The difference between consolidated and unconsolidated conditions was significant for both groups (long consolidation: $b = 1.891$, $z = 19.45$, $p < .001$, short consolidation: $b = 1.275$, $z = 12.62$, $p < .001$). These data support the conclusions drawn from the object recall data.

Lexical decision with novel word primes. Figure 32 (left panel) shows lexical decision RTs to target words when the target was preceded by an associated novel word prime (primed) and when it was preceded by an unassociated novel word prime (unprimed). As before, the RTs were log transformed, and extremely short and long

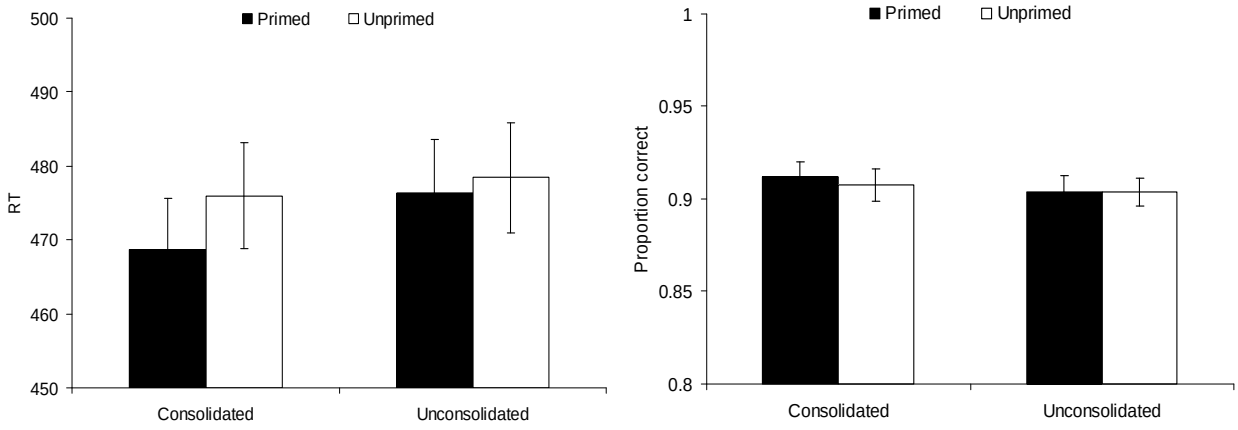


Figure 32. RTs (left panel) and accuracy rates (right panel) in the primed lexical decision task with novel word primes and real word targets. Error bars represent standard error of the means.

RTs were removed (RTs faster than 5 log-ms [148 ms] and slower than 7.3 log-ms [1480 ms]). Only correct responses were included in the RT analysis. A mixed-effects linear model with subjects and items as random factors, and priming (primed vs. unprimed) and time of testing (immediate = unconsolidated, delayed = consolidated, with the long and short consolidation groups combined) as the fixed factors benefitted from subject-specific slopes for trial position. Priming and time of testing did not interact significantly ($p = .10$), hence the interaction was dropped. Averaged over time of testing conditions, there was a significant effect of priming, with faster RTs to primed targets ($b = -0.009$, $t = -2.30$, $p = .03^{\dagger}$). There was an overall RT advantage for consolidated over unconsolidated prime trials ($b = -0.010$, $t = 2.60$, $p = .01$). Visual examination of Figure 32 suggests that there was a priming effect in the consolidated trials, but a much smaller effect in the unconsolidated trials. The lack of an interaction however shows that there was no significant difference in the magnitude of the priming effect between consolidated and unconsolidated conditions, though as the priming effect even in the consolidated condition is small, the lack of an interaction should not be taken as evidence that priming was present in both conditions. To further evaluate the strength of the priming effect in the two conditions, it was assessed separately for both. The priming effect was significant in the consolidated condition ($b = -0.016$, $t = -2.80$, $p = .007$), but was non-significant in the unconsolidated condition. The RT difference between consolidated and unconsolidated conditions was significant only in the primed trials

(b = 0.017, z = 3.00, p = .003). The accuracy rates in this task are displayed in the right panel of Figure 32. A mixed-effects logistic regression model with the same random and fixed factors as in the RT model showed no effect of priming or time of testing, or an interaction between the two.

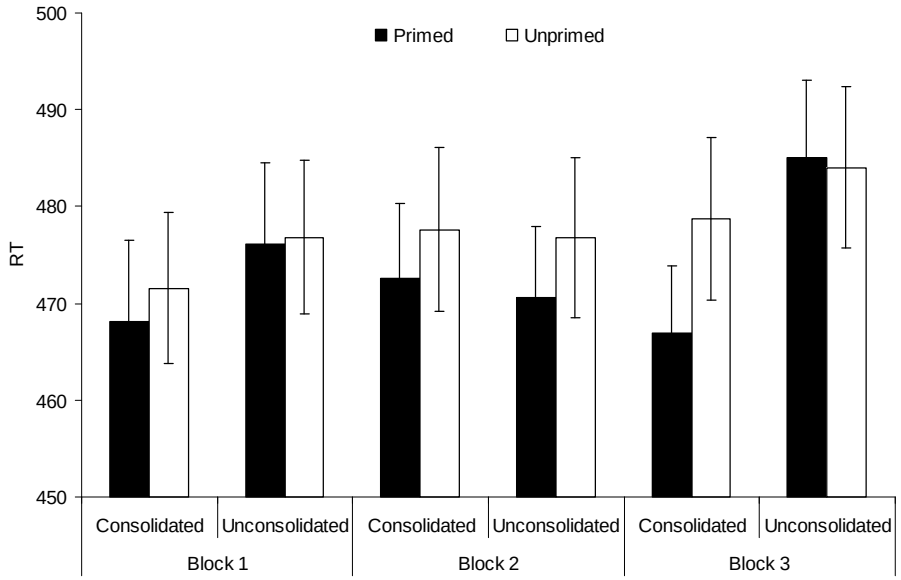


Figure 33. Primed lexical decision RTs in each block with consolidated and unconsolidated novel word primes and real word targets. Error bars represent standard error of the means.

Experiment 4 showed that an advantage for consolidated novel words can emerge over the course of the experiment, when looking at individual blocks of trials. In the present data a similar effect might be reflected in a priming effect emerging only in the later blocks. Hence block was entered as an additional fixed factor to the analysis described above. Figure 33 shows the RT data broken down by block. No significant three-way interaction was found involving priming, time of testing, and block, and was hence dropped. Of the two-way interactions, only the interaction between block and time of testing showed a significant contrast, whereby the difference between RTs to consolidated and unconsolidated trials was higher in block 3 than in block 2 (b = 0.025, t = 2.50, p = .01†). Averaged over priming conditions, RTs did not significantly differ as a function of block in either consolidated or unconsolidated conditions. Although no other interactions involving block reached significance, suggesting that block did not modulate the priming effect in either consolidation condition, the effect of priming was also analysed separately in each block in both consolidation conditions. No priming was found in the

unconsolidated condition in any of the blocks. In the consolidated condition on the other hand priming reached significance in the third block (b = 0.025, t = 2.55, p = .008†). This pattern reinforces the conclusion that no reliable priming was seen in unconsolidated novel words, and that the priming effect that was seen in consolidated words tends to grow stronger over the course of the task, reminiscent of the semantic decision task in Experiment 4.

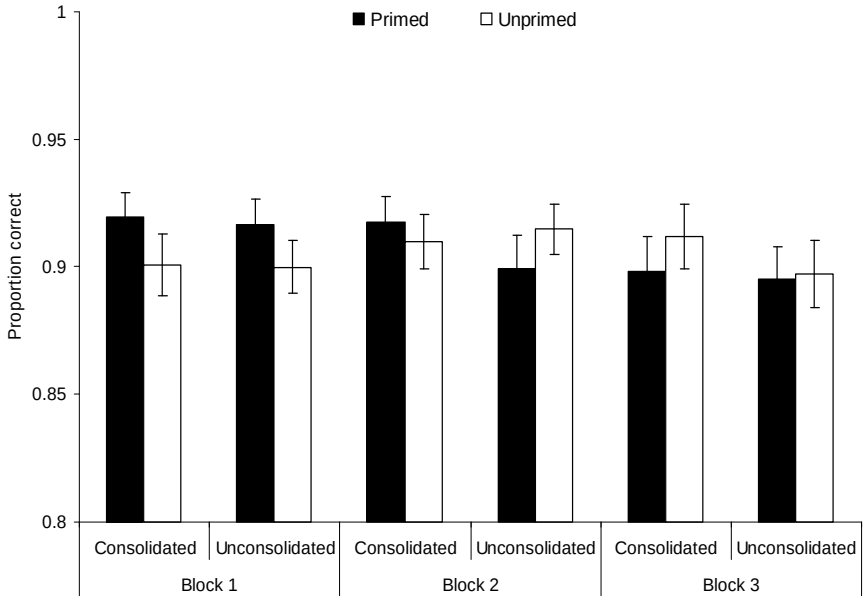


Figure 34. Primed lexical decision accuracy rates in each block with consolidated and unconsolidated novel word primes and real word targets. Error bars represent standard error of the means.

The by-block analysis was carried out for accuracy data as well, by adding block as a fixed factor in the original logistic regression model. These data are presented in Figure 34. No three-way interaction was found. Interaction contrasts involving the effect of priming and block showed that averaged over consolidation conditions there was a significant change in the magnitude of the priming effect from block 1 to block 3 (b = 0.314, z = 1.99, p = .046†), reflecting the initial advantage for primed trials changing into an advantage for unprimed trials. No other two-way interactions showed significant effects. No significant priming effect was found in either time of testing condition in any block though.

Next, the question of whether a long consolidation opportunity results in stronger priming gains than a short consolidation opportunity was assessed. Figure 35 shows RTs in the consolidated condition, broken down by short and long

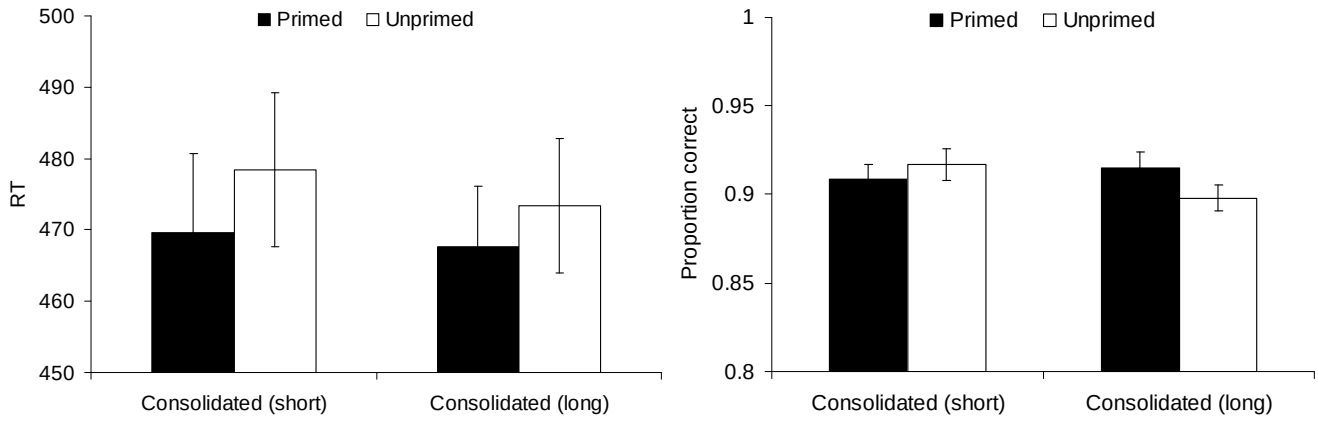


Figure 35. RTs (left panel) and accuracy rates (right panel) in the primed lexical decision task with consolidated novel word primes and real word targets. Short = 1 day consolidation opportunity, long = 1 week consolidation opportunity. Error bars represent standard error of the means.

consolidation opportunity (left panel). A mixed-effects linear model with subjects and items as random factors, and priming (primed vs. unprimed) and length of consolidation opportunity (short or long) as the fixed factors benefitted from subject-specific slopes for trial position. Priming and consolidation length did not interact significantly, hence the interaction was dropped. Averaged over time of consolidation length conditions, there was a significant effect of priming, with faster RTs to primed targets (b = -0.015, t = -2.62, p = .008). There was no overall RT difference between the short and long consolidation conditions. The lack of an interaction shows that there was no statistically significant difference in the magnitude of the priming effect between the short and long consolidation conditions. However, to evaluate the priming effect in closer detail it was assessed in both conditions individually. The priming effect was significant in the short consolidation condition (b = -0.019, t = -2.23, p = .02†), but failed to reach significance in the long consolidation condition (p = .14). There were no significant differences between the two consolidation conditions in either primed or unprimed conditions. Accuracy rates are shown in the right panel of Figure 35. A logistic regression model with the same fixed and random factors as in the RT model was used here. The interaction between priming and consolidation length approached significance (b = 0.343, z = 1.83, p = .067†). There was no significant priming effect in the short consolidation condition, but the effect approached significance in the long consolidation condition (b = 0.240, z = 1.82, p = .069†). No difference between

accuracy was found between the consolidation conditions in either primed or unprimed conditions individually. As the interaction only approached significance, a model without the interaction was also looked at. No significant effect of priming or consolidation length was found in the simplified model.

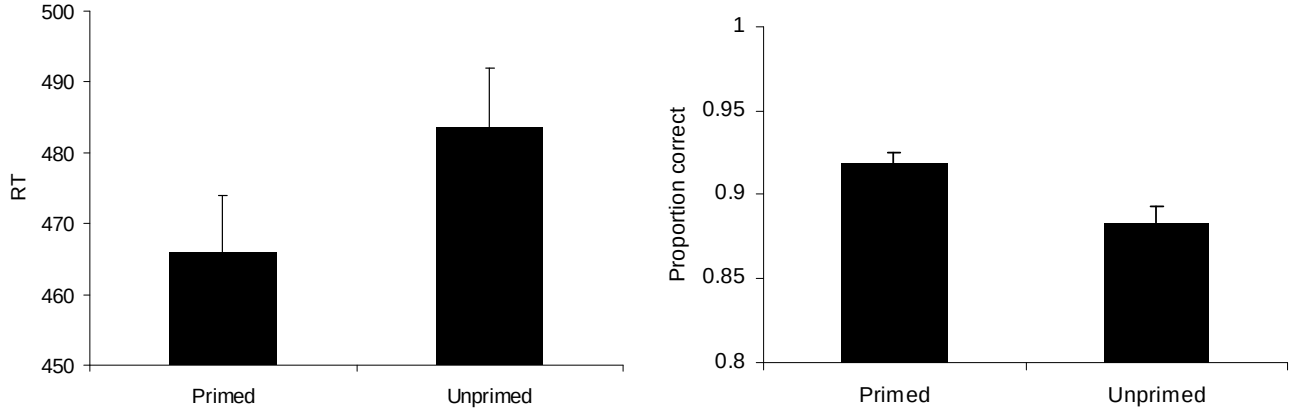


Figure 36. RTs (left panel) and accuracy rates (right panel) in the primed lexical decision task with real word primes and real word targets. Error bars represent standard error of the means.

Lexical decision with real word primes. Figure 36 (left panel) shows lexical decision RTs to real word targets, when the prime was a real word associated (primed) or unassociated (primed) to the target. This real word prime condition was included to make sure the current paradigm results in the typical semantic priming pattern, with faster RTs to primed lexical decision compared to unprimed trials. The data were trimmed in the same way as in the novel word prime condition. A mixed-effects linear model with subjects and items as random factors, and priming (primed vs. unprimed) as the fixed factor showed a significant effect of priming ($b = 0.039$, $t = 7.08$, $p < .001$), with faster RTs to primed trials. To make sure this was the case for both consolidation length groups, consolidation group was added as a fixed factor. Group did not interact with priming, confirming that both participant groups exhibited equivalent priming.

The right panel of Figure 36 shows the accuracy rates for the same task. A mixed-effect logistic regression model with the same random and fixed factors as above showed a significant effect of priming, with fewer errors made to the primed targets ($b = -0.410$, $z = -4.44$, $p < .001$). As above, consolidation length group was added as a fixed factor, but did not enter into an interaction with priming, suggesting that both groups showed a similar priming effect.

Recall that the real word prime condition, like the novel word condition, was divided into three blocks with each prime repeated twice in each block. To evaluate the effect of repetition over block, the priming RT effect was evaluated for each block individually. These data are shown in Figure 37 (left panel). Block was added as a factor to the model described above. No interactions were found between block and the effect of priming. When the interaction was dropped, no significant contrasts involving block averaged over priming were found, showing that RTs overall remained stable across blocks. To make sure the priming effect was equally strong in each block, the effect was evaluated in each block individually. It reached significance in all blocks (block 1: b = 0.034, t = 3.54, p = .001, block 2: b = 0.039, t = 4.07, p < .001, block 3: b = 0.045, t = 4.63, p < .001). RTs did not change over blocks in either priming condition when evaluated individually. Consolidation length group was added as a fixed factor to make sure these conclusions applied equally to both groups. Group did not enter into any interactions with the other factors.

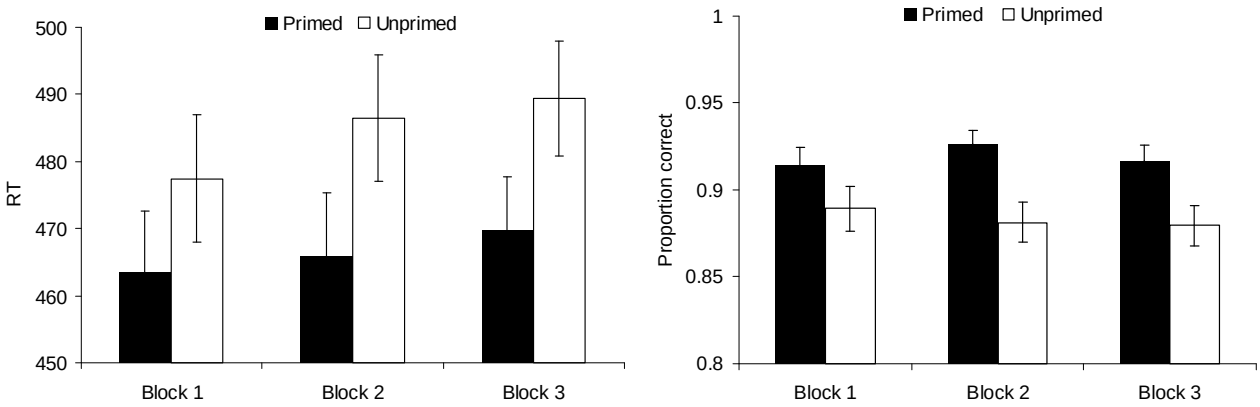


Figure 37. RTs (left panel) and accuracy rates (right panel) in the primed lexical decision task with real word primes and real word targets, broken down by block. Error bars represent standard error of the means.

The right panel of Figure 37 shows the accuracy rates in this task. A mixed-effects logistic regression with the same factors as above showed no interaction between priming and block. When the priming effect was evaluated individually in each block, the effect was not significant in block 1 but reached significance in the other two blocks (block 2: b = 0.565, z = 3.47, p < .001, block 3: b = 0.410, z = 2.60, p = .009). No difference was found between blocks when assessed in primed and unprimed conditions separately. Consolidation length group did not enter into interaction with any of the other variables. These by-block analyses showed that with

real word primes repetition of the primes has no effect on the priming effect over the course of the task, providing an interesting contrast with the novel word prime condition where the effect appeared to be strongest in the last block (although the interaction between priming and block was not significant).

Sentence plausibility judgement. RTs in the sentence plausibility judgement task are shown in Figure 38, left panel, and were analysed using a mixed-effects linear model with subjects and items as random variables, and time of testing (delayed = consolidated, immediate = unconsolidated) and the semantic compatibility of the novel word in the sentence context (match vs. mismatch) as fixed variables. Subject-specific slopes for trial position increased the goodness of fit. Responses faster than 4 log-ms and slower than 7.3 log-ms were removed as extreme scores. No interaction was found between time of testing and sentence compatibility, and the factor was dropped. The simplified model showed a significant difference between the match and mismatch conditions, with faster responses to matching trials ($b = -0.149$, $t = -10.50$, $p < .001$), but no significant difference between consolidated and unconsolidated novel words. To make sure there was no time of testing effect, this effect was evaluated separately for match and mismatch conditions. No difference was found between consolidated and unconsolidated novel words in either compatibility condition. Match vs. mismatch contrasts for each time of testing condition confirmed that the compatibility effect was significant in both consolidated ($b = 0.152$, $t = 7.86$, $p < .001$) and unconsolidated novel words ($b = 0.146$, $t = 7.51$, $p < .001$).

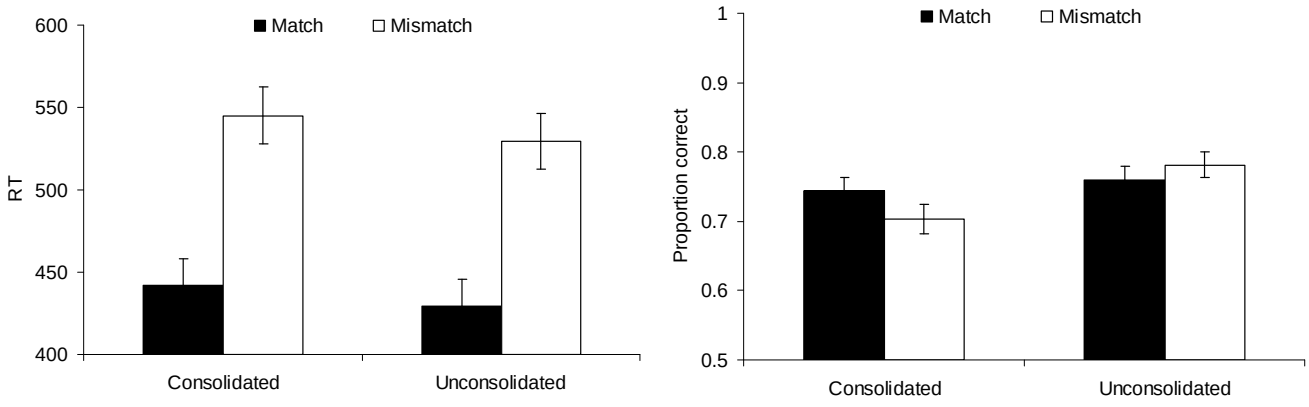


Figure 38. RTs (left panel) and accuracy rates (right panel) in the sentence plausibility judgement task. Error bars represent standard error of the means.

Accuracy rates are shown in the right panel of Figure 38. These were analysed with a mixed-effects logistic regression model with the same random and fixed factors as in the RT analysis. The interaction between time of testing and sentence compatibility was significant, reflecting the larger priming effect in the consolidated over unconsolidated condition (b = 0.313, z = 2.00, p = .046[†]). However, the effect of sentence compatibility did not reach significance when examined for the two time of testing conditions separately. In the mismatch condition there was an accuracy advantage for unconsolidated trials (b = 0.437, z = 4.04, p < .001), which however was not significant in the match condition.

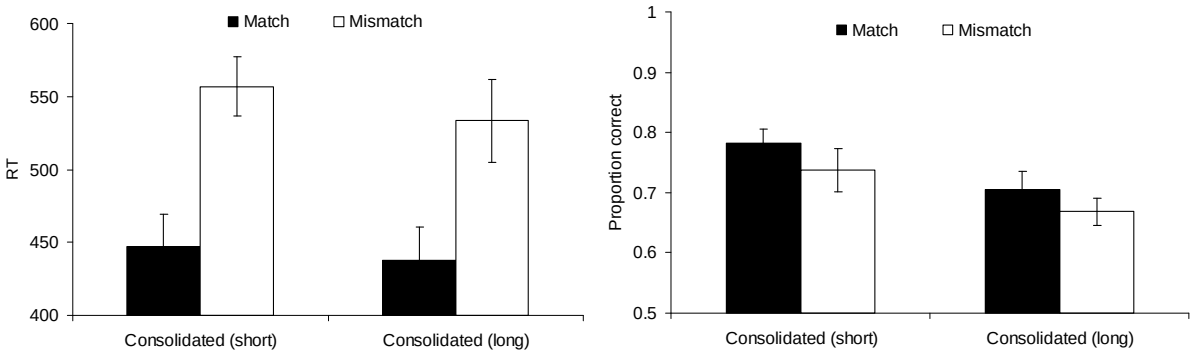


Figure 39. RTs (left panel) and accuracy rates (right panel) in the sentence plausibility task. Short = 1 day consolidation opportunity, long = 1 week consolidation opportunity. Error bars represent standard error of the means.

Figure 39 shows RTs (left panel) for the group who experienced a long consolidation opportunity and for the group who experienced a short consolidation opportunity. A mixed-effects linear model with subjects and items as random factors, and sentence compatibility (match vs. mismatch) and length of consolidation opportunity (short = one day, long = one week) as the fixed factors benefitted from subject-specific slopes for trial position. No significant interaction was found between length of consolidation and sentence compatibility. The simplified model without an interaction showed a significant effect of compatibility (b = 0.146, t = 7.03, p < .001), but no RT difference between the consolidation length groups. As suggested by the lack of an interaction, the compatibility effect was significant in both the short consolidation group (b = 0.173, t = 6.37, p < .001) and the long consolidation group (b = 0.119, t = 4.38, p < .001). There was no RT difference between the groups in either the match or mismatch conditions.

The right panel of Figure 39 shows the corresponding accuracy rates. A mixed-effects logistic regression model with the same fixed and random factors as in the RT analysis was used. There was no significant interaction between the two fixed factors. The simplified model showed no significant effect of sentence compatibility, but did show an overall accuracy advantage for the short consolidation group ($b = 0.368$, $z = 2.21$, $p = .03^{\dagger}$). The compatibility effect failed to reach significance in either consolidation condition individually. When the effect of consolidation length group was examined individually for match and mismatch condition, it reached significance only in the match condition ($b = 0.414$, $z = 2.09$, $p = .04^{\dagger}$).

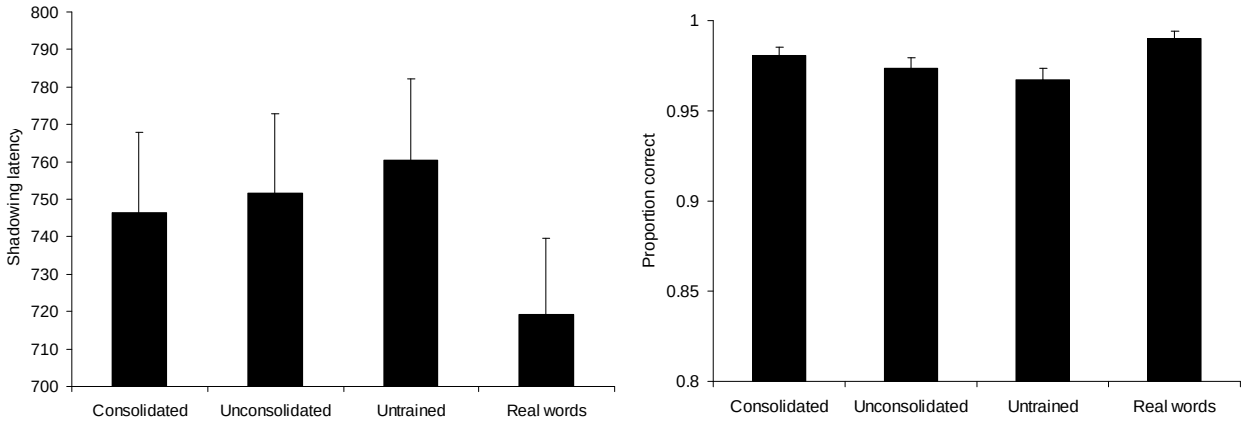


Figure 40. Shadowing latencies (left panel) and accuracy rates (right panel). Error bars represent standard error of the means.

Shadowing. CheckVocal (Protopapas, 2007) was used to check the voice key trigger points, and corrected when necessary. One participant's data in the long consolidation condition were discarded due to a high number of trials lost as a result of recording malfunction. Repetition latencies and accuracy rates are presented in Figure 40. Latencies in log-RTs were analysed using a mixed-effects linear model with subjects and items as random variables, and training condition (consolidated, unconsolidated, untrained, real words) as the fixed variable. Subject-specific slopes for the effect of trial increased the goodness of fit of the model. The same data trimming criteria were used here as in the lexical decision task. Untrained novel words were shadowed slower than any other condition (contrast with consolidated: $b = -0.018$, $t = -4.70$, $p < .001$, unconsolidated: $b = -0.010$, $t = -2.71$, $p = .01^{\dagger}$, real words: $b = -0.054$, $t = -3.98$, $p < .001$). Real words on the other hand were shadowed faster than any other condition (contrast with consolidated: $b = 0.036$, $t = 2.67$,

p = .01†, unconsolidated: b = 0.043, t = 3.22, p = .002). No significant difference was found in shadowing latencies between consolidated and unconsolidated novel words.

Accuracy (Figure 40, right panel) rates in shadowing were analysed next, using a mixed-effects logistic regression model with the same random and fixed factors as in the latency analysis. Accuracy rates to untrained novel words were significantly lower than to consolidated words (b = 0.590, z = 2.72, p = .007), or real words (b = 1.289, z = 4.26, p < .001), but not significantly different from unconsolidated novel words. Real words had higher accuracy rates than either novel word condition (consolidated: b = 0.697, z = 2.18, p = .03†, unconsolidated: b = 1.009, z = 3.26, p = .001). No significant difference was found between the two novel word conditions.

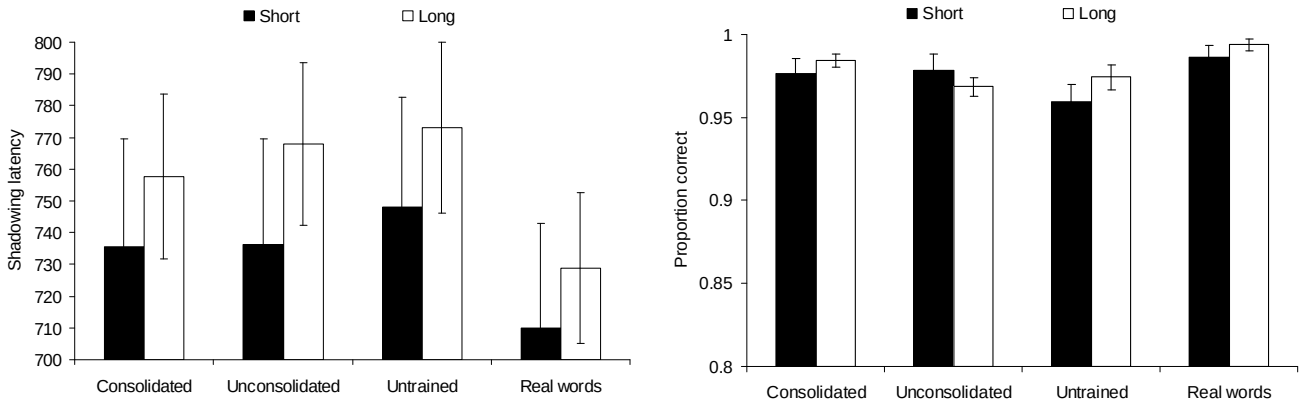


Figure 41. Shadowing latencies (left panel) and accuracy rates (right panel) broken down by length of consolidation opportunity. Error bars represent standard error of the means.

The above analyses found no difference between consolidated and unconsolidated novel words, and appear to replicate the null finding in Experiment 5. However, it is possible that this task too may benefit from a long consolidation opportunity more than a short opportunity. Next, the two consolidation length conditions were analysed by adding length (short or long) as a fixed factor to the model described above. The shadowing latency data are shown in the left panel of Figure 41. A significant interaction was found between training condition and length of consolidation opportunity, whereby the difference between consolidated and unconsolidated conditions was significantly larger in the long consolidation group than in the short group (b = 0.020, t = 2.70, p = .02†). No other interaction contrasts reached significance. Next, the training condition effect was evaluated individually

for both the long and short consolidation groups. In the short consolidation group real words were shadowed faster than any other word condition (contrast with untrained: b = 0.051, t = 3.67, p < .001, unconsolidated: b = 0.037, t = 2.63, p = .02†, consolidated: b = 0.039, t = 2.80, p = .007†). Untrained words were shadowed slower than either unconsolidated or consolidated novel words (b = 0.014, t = 2.77, p = .01†, b = 0.012, t = 2.34, p = .05†). There was no difference between consolidated and unconsolidated conditions. In the long consolidation group real words were again shadowed faster than any other word type (untrained: b = 0.056, t = 4.01, p < .001, unconsolidated: b = 0.051, t = 3.61, p < .001, consolidated: b = 0.033, t = 2.34, p = .03†), while untrained words were slower than consolidated novel words (b = 0.023, t = 4.34, p < .001), but did not differ significantly from unconsolidated novel words. Finally, the difference between consolidated and unconsolidated novel word shadowing times was significant (b = 0.018, t = 3.32, p = .006†) in this group. Visual inspection of Figure 41 suggests that latencies in the long consolidation group were longer than in the short group. Recall however that the length of consolidation condition is a between-groups variable, while the consolidated vs. unconsolidated is within-groups (the error bars in the figures are uncorrected for within-groups comparisons, hence the error bars exaggerate the variability in these contrasts). Statistically the difference between the groups failed to reach significance in each training condition. It did not reach significance when averaged over training conditions either (p = .24).

Accuracy rates are shown in the right panel of Figure 41. Here too a significant interaction was found between training condition and length of consolidation opportunity, whereby the difference between unconsolidated and untrained conditions was significantly different in the short consolidation group than in the long consolidation group, reflecting the reversal of the effect between the groups (b = 0.919, z = 2.26, p = .02†). Another contrast that was marginally significant showed that the difference between consolidated novel words and real words was larger in the short consolidation group than in the long consolidation group (b = 1.184, z = 1.97, p = .05†). Next, the differences between training conditions were evaluated separately for the two consolidation opportunity groups. In the short opportunity group real words differed significantly only from the untrained words (b = 1.211. z = 3.36, p < .001). Untrained words on the other hand differed from all conditions (unconsolidated: b = 0.740, z = 2.56, p = .01†,

consolidated: b = 0.625, z = 2.23, p = .03[†]). There was no difference between consolidated and unconsolidated novel words. In the long opportunity group real words had higher accuracies than either untrained words (b = 1.463, z = 2.94, p = .003[†]) or unconsolidated novel words (b = 1.642, z = 3.34, p < .001). Untrained words did not differ from either novel word condition. Consolidated and unconsolidated novel words on the other hand did show a significant difference (b = 0.740, z = 2.22, p = .03[†]). There was no difference between long and short consolidation groups in each training condition, or when averaged over all training conditions. Hence both latency and accuracy data show that shadowing of novel words did benefit from consolidation, but only if a consolidation opportunity of one week was offered.

5.2.3 Discussion

The primary aim of Experiment 6 was to evaluate the degree to which newly learned words afford semantic priming in a task which was calibrated to rely on strategic semantic processing more than automatic processing. The semantic priming task using real word primes confirmed that the task parameters were such that a robust priming effect of 18 ms was found. This was also reflected in accuracy rates, with higher accuracies for primed lexical decisions. Numerically 18 ms is a fairly small semantic priming effect. It is important to consider why this might be the case. The main reason is likely to be the constraints in choosing the stimuli. Due to the repetition of primes in both the real and novel word conditions, only primes with three strong associates could be used. For most words it is impossible to find three very strong associates, hence the overall association strengths in the present experiment are lower than would be the case if only the very strongest prime-target pairs could be chosen (which is the case in the priming research literature in general).

The critical condition using novel word primes also showed a priming effect. This effect was significant when averaged across consolidation conditions. While there was no statistically reliable difference in the magnitude of the priming effect using consolidated (7 ms) and unconsolidated (2 ms) novel words, the effect was significant only in consolidated novel word primes when the two conditions were evaluated separately. This suggests that novel word meanings do benefit to some

degree from offline consolidation. The comparison between a short consolidation opportunity of one day and a long consolidation opportunity of one week showed no significant difference, the priming effect was 9 ms in the short condition and 6 ms in the long condition, suggesting that priming emerges within the first 24 hours (possibly associated with sleep) and does not significantly grow over the next seven days. These effects were again numerically small, but reasonable when comparing them to the real word prime condition which too was quite small.

Looking at the development of the semantic priming effect as the lexical decision task progressed revealed a similar pattern as seen in Experiment 4. Recall that in Experiment 4 a consolidation effect in semantic decision only emerged in the last block of the task, in the form of faster RTs to trials with consolidated as opposed to unconsolidated novel words. In the present experiment also the priming effect in consolidated novel words reached significance only in the third block of the task (although the interaction between block and priming was non-significant). It is important to note however that there was a numerical effect in the first two blocks as well, while in the semantic decision task of Experiment 4 there was no hint of a consolidation effect prior to the third block. Nonetheless, these data perhaps reflect a process whereby the fragile consolidation effect becomes detectable only after participants have had a chance to access the early learned consolidated meanings a few times. It may be that this process is needed to overcome an initial advantage for the recently learned unconsolidated novel words which may benefit from an episodic recency effect. This line of reasoning is further supported by the statistical lack of an effect of block in the real word prime condition (although numerically the effect seemed to grow stronger from block 2 onwards), suggesting that the block effect may be specific to novel word meaning access.

The semantic consolidation effect is particularly intriguing in light of the explicit meaning recall data. Here we saw an advantage for the recently learned novel words. The same pattern was described in Chapter 4. Furthermore, a comparison of the short and long consolidation groups showed that the long consolidation group appeared to have forgotten more consolidated meanings than the short consolidation group. Here we see a striking dissociation between explicit recall of novel word meanings, which seems to be subject to decay as a function of time passing, and an online measure of speed of novel word meaning access, which seems to benefit from passing time, at least during the first 24 hours. This dissociation

supports the notion that semantic priming and explicit meaning recall here measured two fundamentally different types of semantic access. However, it is important to note that the apparent decline in explicit meaning recall may also be at least partially caused by interference from having to learn a second set of novel words before testing. Experiment 7 provided an opportunity to test this alternative hypothesis.

The RSVP version of the sentence plausibility task showed no effect of consolidation. Responses were faster to novel words that were semantically congruent with the sentence context, and this was the case for both consolidated and unconsolidated novel words. There was no significant difference in the magnitude of the effect between the consolidation conditions, and the effect was nearly identical irrespective of the length of the consolidation opportunity. This is interesting because in Experiment 4 there was a consolidation effect whereby the semantic congruency effect was significantly larger in consolidated novel words. However, in that experiment the reading of the sentence was self-paced, with participants allowed to first read the sentence at their own pace and then press a key to reveal the novel word. As discussed in the previous chapter, this procedure allows the use of guessing strategies, where participants may have generated hypotheses about the identity of the novel word before seeing it, in which case upon revealing the novel word they would merely need to check if the given word matched the expected word. Such a process is likely to involve orthographic influences more than semantic processing, with learning of word forms having already been shown to benefit from consolidation (c.f. cued recall in Experiment 5). The current version of the task did not allow participants time to make explicit guesses about the novel word, and required a speeded response about the identity of the novel word at the end of the sentence presentation. This makes it more comparable to the semantic decision task of Experiment 4 rather than the priming task of the present experiment. Consequently it is possible that the sentence task would have required a larger number of trials and repetitions for any consolidation effect to emerge.

Finally, the shadowing task provided intriguing data. When averaged over the two consolidation length conditions, there was no evidence of a consolidation benefit in shadowing latencies between consolidated and unconsolidated novel words. However, when the two consolidation length groups were examined separately, a robust consolidation effect was seen in the long consolidation group but not in the short consolidation group. This may resolve the discrepancy between Experiment 4

where consolidation in shadowing was seen, and Experiment 5 where it was not seen. Shadowing may be a task that benefits from several days or nights of consolidation. Experiment 4 suggested that a consolidation benefit may be observed after a one day delay, but judging by Experiment 5 the effect is not robust enough to be observed consistently. The contrast here between short and long consolidation suggests that the consolidation process continues over several days, suggesting it may be a slow, incremental process. In such circumstances it is to be expected that the effect is robust enough to be detected within 24 hours only occasionally. Whether or not it is seen after such a short delay may depend on a number of factors, such as the exact timing between the training and testing, and possibly also depending on individual differences between participants' sleeping patterns or memory proficiency. Future studies looking at shadowing and consolidation would need to control or manipulate these types of factors more carefully. Participants in the long consolidation group tended to shadow slower than the short group (although this difference did not reach significance). This was probably due to the increased delay between training and test in the long group. I will return to some of the properties of the shadowing task that may explain its reliance on long term consolidation in the General Discussion.

In sum, the current experiment showed that semantic priming using novel words benefits from offline consolidation within the first 24 hours, at least when strategic semantic access is the primary source of priming. The next experiment will focus instead on automatic semantic activation.

5.4 Experiment 7

Experiment 6 showed in a semantic priming task with visible primes that newly learned meaningful words can prime lexical decision, but that such priming effects are not reliably seen until one day after training. While the priming task used in that experiment required fast online access to the novel word meanings, it cannot be said to measure purely automatic semantic access, as the prime was clearly visible, and participants were aware of the semantic relationship between prime and target. In fact, the task parameters (long SOA, visible prime) were intentionally tuned so that while both strategic and automatic processes were likely to contribute to the priming effect, the contribution of strategic processes was probably larger.

The main aim of Experiment 7 was to examine novel word semantic priming in a task where automatic semantic activation was the primary source of priming and strategic effects were minimised. For this purpose I chose to use the three-field masked priming paradigm introduced by Forster and Davis (1984). In this paradigm a briefly presented prime is immediately replaced (and masked) by the target, resulting in a very short SOA. Furthermore, the prime is also masked by preceding presentation of a visual mask (e.g., "##########"). Several authors have reported semantic priming effects using this methodology, both with associated and semantically related prime-target pairs, and with a range of short SOAs (e.g., Sereno, 1991; Perea & Godor, 1997; Rastle, Davis, Marslen-Wilson, & Tyler, 2000; Bueno & Frenck-Mestre, 2008). Although masked repetition and form priming have been used with novel words (e.g., Forster, 1985, Qiao et al., 2009), to my knowledge masked semantic priming has not been examined in newly learned words. Interestingly though, masked semantic or associative priming effects have been reported between languages (e.g., Dutch prime – English target) in bilingual participants. Gollan, Forster, and Frost (1997) showed masked translation priming (priming between the same two words in different languages) between Hebrew-English cognates (words similar in meaning and form) and noncognates (words similar in meaning only), although only with L1 primes. Perea, Dunabeitia, and Carreiras (2008) looked at priming of different but related/associated words in two different languages (Spanish and Basque), and found equivalent masked priming both between and within languages.

The extremely short SOA in masked priming experiments is argued to reduce strategic effects, as participants do not have time to generate expectations before seeing the target. One line of evidence for this claim comes from the manipulation of RP in priming experiments using masked and visible prime conditions. Recall that RP manipulation is thought to give rise to strategic effects, hence a task not sensitive to RP manipulation is likely to be less affected by strategic factors. Although Bodner and Masson (2003) did find a larger masked priming effect in a high RP condition than in a low RP condition, Grossi (2006) failed to replicate this effect and instead saw similar masked priming in high and low RP conditions. Furthermore, using the same materials, Grossi (2006) did report an RP effect but only when using visible primes. The lack of an RP effect has also been reported by Perea and Rosa (2002),

further supporting the claim that masked priming relies primarily on automatic semantic access.

One of the most contentious issues in the masked priming literature is the question of whether semantic priming can take place outside of consciousness. The debate here revolves around the issue of under what circumstances can a prime be considered to have been presented subliminally (for an exhaustive review, see Kouider & Dehaene, 2007, and for a recent meta-anlysis, see Van den Bussche, Van den Noortgate, & Reynvoet, 2009). This is however an issue that most researchers interested in language processing have not commented on, in fact many experiments looking at semantic priming under masked conditions have not even attempted to establish whether participants were aware of the primes or not. This is probably because the short SOA and the increased difficulty in perceiving the prime as a result of very brief prime duration have been considered to be sufficient conditions for minimising the contribution of strategic processes. This is also the position adopted in the present approach.

A second aim of Experiment 7 was to re-examine the advantage observed for recently learned word meanings in the explicit meaning recall task in the experiments reported in the previous chapters and in Experiment 6. In these experiments it consistently appeared that participants' recall of word meanings learned a day or a week earlier had decreased, and was worse than recall of recently learned meanings. Because in those experiments recall was always tested after the second training session, it is possible that the cost associated with the earlier learned words was due to interference from the recent learning of novel materials. To see if this was the case, Experiment 7 used only one set of trained novel words, and tracked the recall of the word meanings as a function of time.

Three testing times were included: immediate test after training, test on the following day, and a third test one week after training. It is possible that when looking at consolidation of semantic information, more than 24 hours is needed for observable effects to emerge. This may be the case particularly when looking at automatic semantic access. For example, Clay et al. (2007) only observed a semantic picture-word interference effect using novel words after one week from training (although these authors did not include testing sessions earlier, apart from the immediate test). Dagenbach et al. (1990) found semantic priming with novel primes

only after five weeks of training. Hence it was deemed necessary to re-test after a week in case the effect requires several days to emerge.

5.4.1 Method

Materials

The materials were drawn from the stimulus pool used in Experiment 6. However, since only one set of words was trained in the current experiment, only 34 novel words and 34 elaborated meanings were needed. Novel words was selected so that the full range of word lengths from the original set was represented (M = 6.4 letters, range = 5-8), and that the words were as dissimilar as possible from each other to minimise their confusability. A set of 34 meanings and corresponding associates was selected on the basis of the accuracy data from Experiment 4. Recall that in that experiment participants were asked to judge whether a novel word meaning and a real word target were related. This enabled the selection of those 34 meanings with their corresponding three associates that resulted in the highest accuracy rates in the semantic decision task, suggesting that for this participant population these word-associate pairs were the most likely ones to generate semantic priming. The mean CELEX frequency of the new set of nouns representing the meaning objects was 49.4, and the mean frequency of their associates (real word targets in lexical decision) was 99.6, mean length in letters was 6.0, and mean association strength between the objects and the targets was 0.18. Nonword targets were derived from the real word targets, hence they were matched to the real word targets in all respects (see Experiment 6). The novel words and meanings used in this experiment are indicated in Appendices 4 and 5.

The same set of 34 real word primes and their three associated targets used in the real prime condition of Experiments 4 and 6 was used here. The properties of these stimuli are described in Chapter 4. Nonword targets derived from the selected real word targets for use in the lexical decision task were taken from Experiment 6.

Each novel word was paired with a meaning, taking care not to pair meanings with words that sounded anything like the novel word. The same pairings were used for all participants.

Design

The experiment was carried out on three days (Figure 42), spanning one week. On day 1, participants were trained on the novel words. This was immediately followed by a testing session. For about half of the participants (n = 21) a testing session consisted of the explicit meaning recall task, followed by primed lexical decision, and an identical testing session was attended on day 2 and day 8. For the remaining participants (n = 27) the testing session on day 1 and day 2 included only the primed lexical decision task, and the day 8 session included the lexical decision task followed by a meaning recall task. This allowed me to see if the repeated administration of the meaning recall task is needed to maintain memory of the word meanings, with potential consequences for explicit recall and semantic priming. This would appear to be the case if priming were found only in the group of participants who experienced the meaning recall task in the beginning of each testing session.

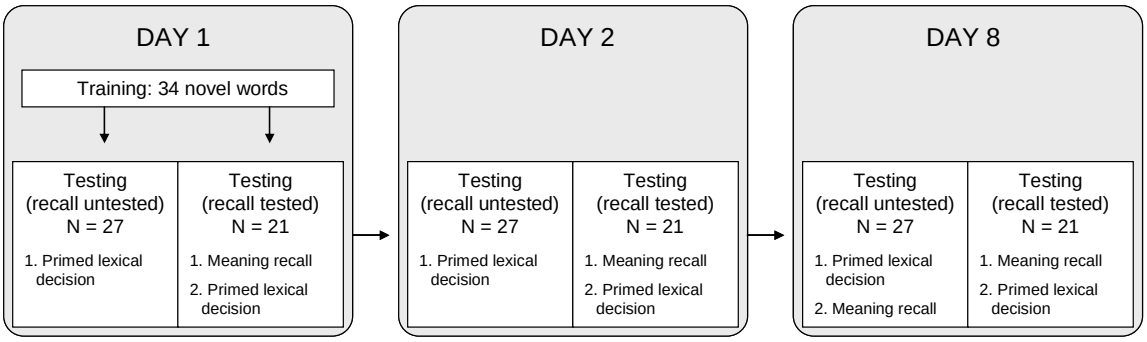


Figure 42. Timing of training and testing sessions in Experiment 7.

For the purposes of the primed lexical decision task, a split-plot design was used as before, and the same counterbalancing measures were taken as in Experiment 6, to ensure that all targets appeared in primed and unprimed conditions across participants, while each individual participant saw each target only once. As in Experiment 6, the task consisted of three blocks with each prime repeated twice in each block, once with a word target and once with a nonword target. The order of blocks using real word primes and novel word primes were counterbalanced as before so that half of the participants did the blocks using real word primes first, and vice versa. Note that due to experimenter error two novel word primes were removed from the lexical decision analyses because they were paired with wrong targets. Hence the primed lexical decision data is based on 32 novel words, but analyses regarding meaning recall include the full set of 34 novel words.

Procedure

Training. The procedure of the training session was identical to that used in Experiment 6, with meaning-to-word matching, word-to-meaning matching, meaning recall, and semantic plausibility tasks. The number of exposures was also the same as before (17 in total). E-prime was used for stimulus presentation both in training and testing and the same computer equipment was used as in Experiment 6.

Testing. Meaning recall was identical to Experiment 6, with a novel word presented on screen and responses typed on the keyboard, without a time limit or feedback.

The order of trials in the primed lexical decision task was pseudorandomised individually for each participant in each testing session using the same constraints and software as in Experiment 6. A trial started with the presentation of a mask (##########) for 500 ms in the centre of the screen. The number of #s was the same on each trial, and was determined by the length of the longest stimulus used in the experiment (i.e., ten letters). The prime word in lowercase letters appeared at the offset of the mask for 47 ms, and was then replaced by the target presented in uppercase letters. The target remained on screen until a response was made, or until 2000 ms had elapsed. At this point accuracy and RT feedback was presented. A key press initiated a new trial with a delay of 500 ms. Participants were not told about the existence of the prime, and were instructed to make the lexical decision as quickly and as accurately as possible by pressing a key on a Cedrus button box labelled "Word" or "Nonword". Half of the participants responded "word" using their right hand, and the other half used the opposite mapping. At the end of the last testing session participants were asked if they had noticed the prime word or anything else happening on the screen between the presentation of the mask and the target. Out of 48 participants only seven reported noticing that a word would sometimes appear briefly, but none reported being able to read the word. Although this does not mean that the primes were fully outside of consciousness (e.g., Kouider & Dupoux, 2004), it does suggest minimal opportunity to use of strategic responding.

Participants

Forty-eight students from the University of York took part in the experiment (8 male, 9 left-handed, mean age = 20.0, range = 18-40). No participants reported

language disorders, or had participated in Experiments 4-6. Participants were paid or received course credit, and the most accurate and fastest 50% of the participants were entered into a prize draw for a £10 gift certificate.

5.4.2 Results

Training data

Accuracy rates were analysed in the training task on day 1. These data are presented in Figure 43 (objects on the left y-axis, features on the right y-axis). The object data were analysed first. A mixed-effects logistic regression model with subjects and items as random factors, and block (block 1, block 2, block 3) and testing group (explicit recall tested in each session = tested, or tested only in the end of the experiment = untested) as the fixed factor was fitted. Subject-specific slopes for the effect of block improved the goodness of fit. Testing group did not show any significant contrasts confirming that there was no difference between the two groups at training. Hence this variable was dropped. Accuracy rates increased significantly from block 1 to block 2 ($b = 1.818$, $z = 15.26$, $p < .001$), from block 1 to block 3 ($b = 3.104$, $z = 18.25$, $p < .001$) and from block 2 to block 3 ($b = 1.283$, $z = 9.77$, $p < .001$). Feature recall was analysed next using ordinal logistic regression with block (block 2, block 3) as the fixed factor. Number of features recalled increased

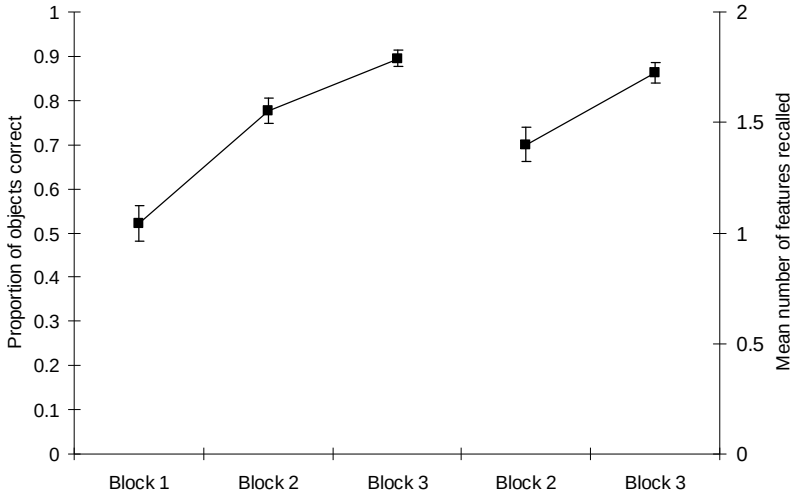


Figure 43. Accuracy rates in the meaning recall training task. Error bars represent standard error of the means.

significantly from block 2 to block 3 (b = 1.023, z = 12.14, p < .001). Accuracy rate in the sentence plausibility training task was very high, with proportion of correct responses at 0.96.

Testing data

Meaning recall. Figure 44 shows the accuracy data in the meaning recall test task for proportion of objects recalled (left panel) and mean number of features recalled (right panel). These data are shown both for the group who were tested in the beginning of each test session, and for the group which was tested only once, in the end of the last test session. Data from the repeatedly tested group were analysed first. A mixed-effects logistic regression model with subjects and items as random factors, and time of testing (day 1, day 2, day 8) as the fixed factor was fitted. Subject-specific slopes for the effect of time of testing were added. While no significant difference was seen between day 1 and day 2, accuracy rates on day 8 were significantly lower than either on day 1 (b = -1.697, z = -7.62, p < .001) or day 2 (b = -1.439, z = -6.96, p < .001). Next, the difference between the repeatedly tested and once only tested groups on day 8 was compared, using an identical model as above but adding testing group as a fixed factor and removing block. A significant difference was found between the groups, with the repeatedly tested group recalling more novel word objects (b = 0.982, z = 2.77, p = .006).

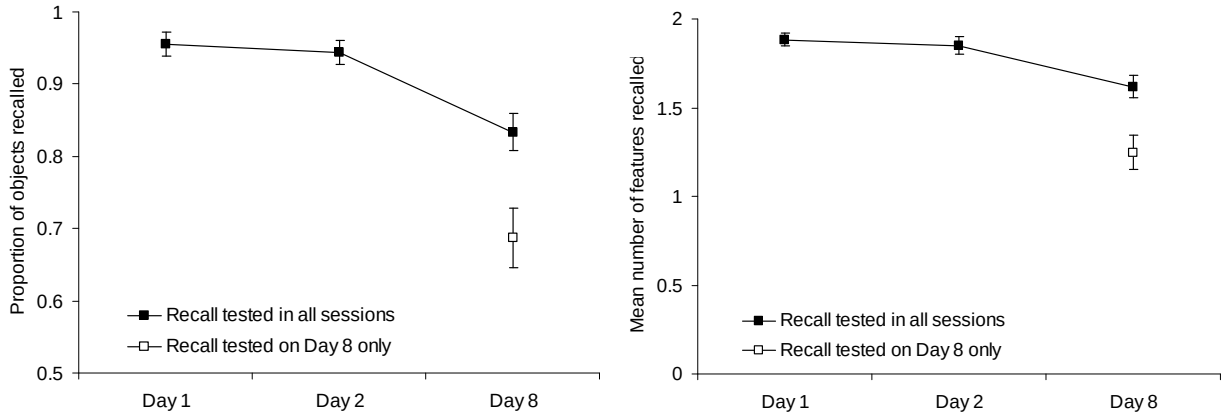


Figure 44. Accuracy rates in the meaning recall test task. Error bars represent standard error of the means.

Next, the number of features recalled in the repeatedly tested group was analysed using an ordinal logistic regression model with time of testing as the fixed

factor (right panel of Figure 44). Again, no difference was found between day 1 and day 2, but recall rates were significantly lower on day 8 than on day 1 (b = -1.874, z = -12.13, p < .001) or day 2 (b = -1.594, z = -11.44, p < .001). A comparison of features recalled between the repeatedly tested and once only tested groups showed a significant difference, with the repeatedly tested group recalling more features (b = 1.639, z = 17.58, p < .001). It seems then that when participants learn only one set of novel words, there is no significant decline in recall rates after one day. A decline is seen on the other hand one week after training. Also, the low recall of the non-tested group suggests that repeated testing helps maintain higher explicit recall rates than in the absence of testing.

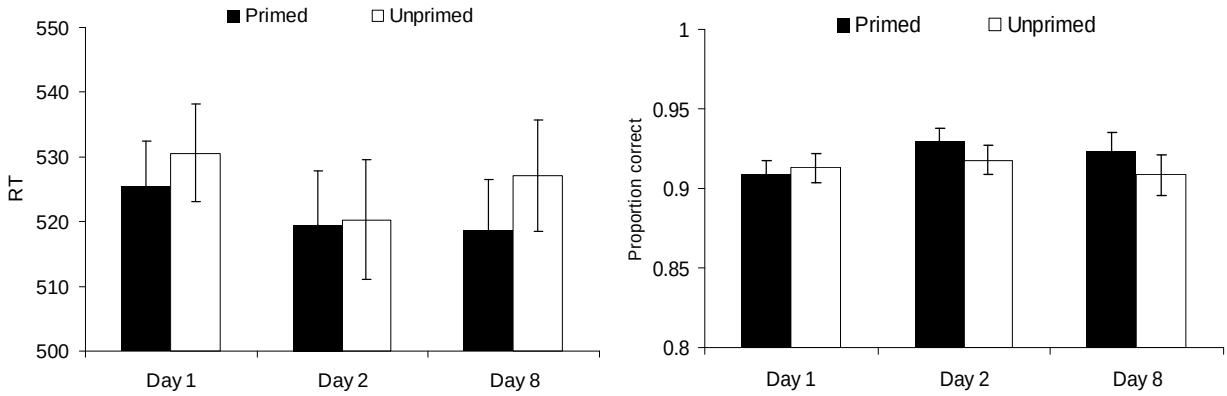


Figure 45. RTs and accuracy rates in the primed lexical decision task with novel word primes and real word targets. Error bars represent standard error of the means.

Lexical decision with novel word primes. The left panel of Figure 45 shows lexical decision RTs to target words preceded by semantically associated novel words (primed) and semantically unassociated novel words (unprimed). The data were again log transformed and extremely fast and slow responses were removed (RTs faster than 5 log-ms [148 ms] and slower than 7.3 log-ms [1480 ms]). Only correct responses were included in the RT analysis. A mixed-effects linear model with subjects and items as random factors, and priming (primed vs. unprimed) and time of testing (day 1, day 2, day 8) as the fixed factors benefitted from subject-specific slopes for trial position, and item-specific slopes for counterbalancing list. A significant interaction contrast showed that the priming effect was larger on day 8 than on day 2 (b = 0.015, t = 2.14, p = .03†). In fact, when the priming effect was assessed on each day individually, it reached significance on day 8 (b = 0.015,

t = 2.71, p = .007[†]), but was not significant on day 2 or day 1. Time of testing also affected RTs in that RTs to both primed and unprimed conditions became significantly faster from day 1 to day 2 (primed: b = -0.015, t = -3.15, p = .003, unprimed: b = -0.023, t = -4.79, p < .001). Primed responses also became faster from day 1 to day 8 (b = -0.015, t = -3.16, p = .001), but not from day 2 to day 8. Unprimed responses became slower from day 2 to day 8 (b = 0.015, t = 3.01, p = .002), and did not show any change from day 1 to day 8.

Accuracy rates in this task are shown in the right panel of Figure 45, and were analysed with a mixed-effects logistic regression model with the same random and fixed factors as in the RT analysis. A marginally significant interaction between time of testing and priming was found, whereby the priming effect was larger on day 8 compared to day 1 (b = 0.298, z = 1.89, p = .06[†]). The priming effect was not significant on day 1 or day 2, but did reach significance on day 8 (b = 0.238, z = 2.08, p = .04[†]). Accuracy to primed trials increased from day 1 to day 2 (b = 0.324, z = 2.82, p = .005[†]) and from day 1 to day 8 (b = 0.242, z = 2.14, p = .03[†]). Accuracy to unprimed trials did not change as a function of day.

The analysis on the full set of participants suggested that the masked priming effect emerged reliably only in the last testing session, on day 8. The next analysis attempted to establish whether this pattern was seen in both those participants whose explicit recall of the novel word meanings was tested in the beginning of each session (tested group) and in those participants whose recall was only tested at the end of the last testing session (untested group). The RT data for each of these groups are displayed in Figure 46 (upper panel). These data were analysed by adding testing group (tested vs. untested) to the model used above to analyse the RTs. No three-way interactions reached significance, thus the term was discarded. Of the two-way interactions involving test group, interactions between the effect of time of testing and test group showed that the difference between day 1 and day 2 RTs was larger for the untested group than the tested group (b = 0.028, t = 4.03, p < .001). The same was true of the difference between day 2 and day 8 (b = 0.020, t = 2.91, p = .004[†]). Looking at the effect of time of testing on RTs in the untested group, the difference between day 1 and day 2 was significant (b = 0.031, t = 6.88, p < .001), as was the difference between day 1 and day 8 (b = 0.015, t = 3.37, p < .001), and between day 2 and day 8 (b = 0.016, t = 3.51, p < .001). Time of testing did not modulate RTs in the tested group. The RT difference between the tested and untested groups however

failed to reach significance on each day. The most interesting contrasts however involve the priming effect. While the lack of an interaction between test group and priming suggests that both groups showed the same priming effect, the effect was also evaluated for both groups on each day. The untested group showed no priming effect on days 1 and 2, but did show a significant effect on day 8 ($b = 0.017$, $t = 2.38$, $p = .02^{\dagger}$). In the tested group the priming effect failed to reach significance on all days. Note however that this was likely to be due to reduced statistical power. It is worth pointing out also that at least numerically the data in Figure 46 suggest that the trend towards a priming effect on day 1 seems to have been carried by the tested group only.

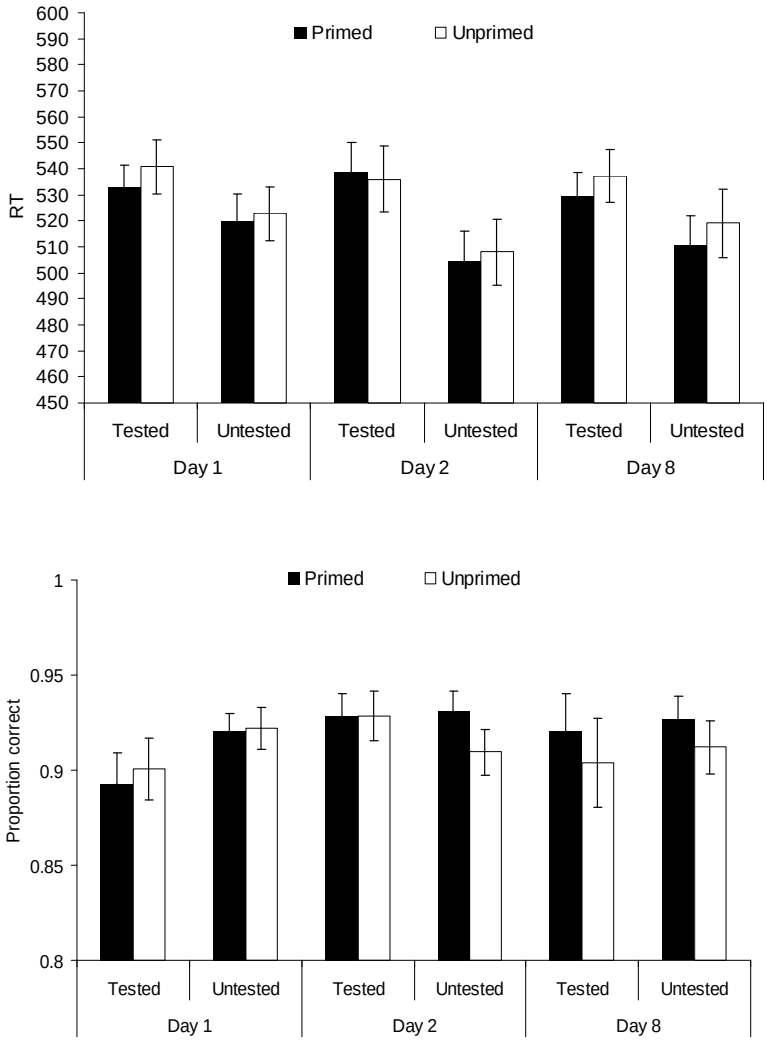


Figure 46. RTs and accuracy rates in primed lexical decision with novel word primes, broken down by testing condition. Error bars represent standard error of the means.

Accuracy data were also analysed by adding testing group as a fixed factor (Figure 46, lower panel). No three-way interactions reached significance. Of the two-way interactions, contrasts only involving testing group and day of testing reached significance, whereby the tested group improved from day 1 to day 2 significantly more than the untested group ($b = 0.451$, $z = 2.78$, $p = .005^{\dagger}$). The tested group improved significantly from day 1 to day 2 ($b = 0.445$, $z = 3.70$, $p < .001$), and from day 1 to day 8 ($b = 0.231$, $z = 2.00$, $p = .045^{\dagger}$). Recall accuracy in the untested group did not change as a function of day. No significant difference was found between the tested and untested groups on any day. Priming did not interact with testing group. However, looking at the effect of priming in each testing group individually on each day, the priming effect reached significance only in the untested group, on day 2 ($b = 0.342$, $z = 2.21$, $p = .03^{\dagger}$).

Next, the effect of priming was evaluated in each testing group in each of the three blocks. The RTs for the tested group are shown in the upper panel of Figure 47. A mixed-effects linear model with subjects and items as random factors and priming (primed vs. unprimed), time of testing (day 1, day 2, day 8) and block (block 1, block 2, block 3) was fitted on these data. Trial position and counterbalancing list were again included in subject and item-specific slopes. No three-way interactions reached significance. Only one two-way interaction contrast reached significance, showing that the priming effect in block 1 on day 1 changed significantly on day 2, reflecting the reversal of the effect. The priming effect in this group reached significance only in block 1 of day 1 ($b = 0.028$, $t = 2.16$, $p = .03^{\dagger}$). No other effects involving the other factors reached significance. RTs for the untested group are shown in the lower panel of Figure 47. In this analysis no three-way interactions reached significance. Of the two-way interactions the only significant contrast involved block and time of testing, whereby the contrast between day 1 and day 2 was larger in block 3 than in block 1 ($b = 0.024$, $t = 2.13$, $p = .03^{\dagger}$). Looking at the effect of block on RTs averaged over primed and unprimed trials, the only significant contrast was found on day 2, between block 1 and block 3 ($b = 0.031$, $t = 2.17$, $p = .03^{\dagger}$). The lack of an interaction between block and priming suggested block did not modulate priming, and this was further confirmed by the absence of a priming effect in each block on each day. It is again worth noting that the lack of significant priming effects here is likely due to reduced power, as the number of both participants and number of data points per participant is greatly reduced from the

more powerful analysis described earlier. It is also possibly important that numerically the priming effect on day 1 seems to be strongest in the first block and become attenuated in the second and third blocks.

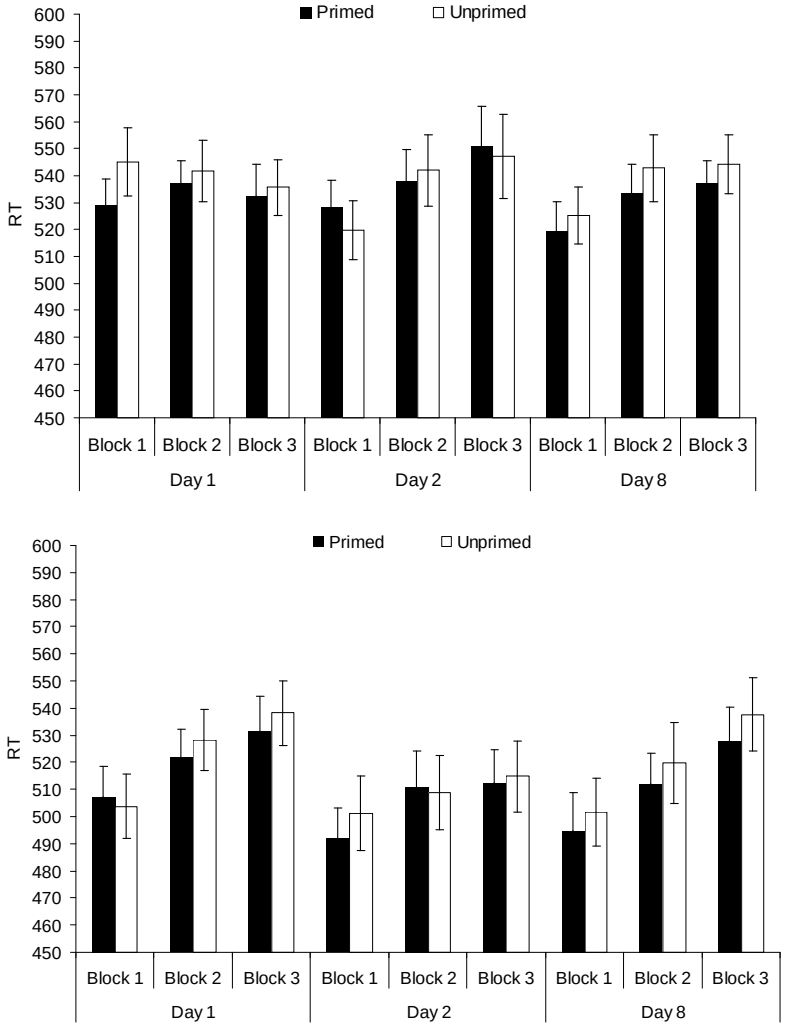


Figure 47. RTs in primed lexical decision with novel word primes, broken down by block and testing condition (upper panel = tested, lower panel = untested). Error bars represent standard error of the means.

Accuracy rates were similarly analysed using a logistic regression model with the same factors as in the RT model. Data from the tested group are presented in the upper panel of Figure 48. A logistic regression model with the same random and fixed factors as above was used. No three-way interactions were found, and the term was dropped. No two-way interactions reached significance either. On day 1 responses in block 3 attracted more errors than block 1 ($b = -0.449$, $z = -2.01$, $p = .04^{\dagger}$). The priming effect was significant only in the first block of day 8

(b = -0.609, z = -1.98, p = .048†). The reverse priming effect on day 1, block 3 was marginally significant (b = 0.485, z = 1.87, p = .06†).

The same analysis was carried out for the untested group (Figure 48, lower panel). No three-way interactions reached significance. Of the two-way interactions, priming interacted with block, in that on day 1 the priming effect was significantly different in block 3 compared to block 1 (b = 0.469, z = 2.14, p = .03†) and block 2 (b = 0.433, z = 2.03, p = .04†). The same was true on day 2 and day 8 where the priming effect on block 1 was attenuated on block 3 (day 2: b = 0.467, z = 2.13, p = .03†, day 8: b = 0.469, z = 2.13, p = .03†). On day 8 the effect on block 2 was also attenuated compared to block 3 (b = 0.432, z = 2.02, p = .04†). The priming effect was marginally significant on day 2 in block 1 (b = -0.541, z = -1.96, p = .05†), and significant on day 8 in block 2 (b = -0.663, z = -2.39, p = .02†).

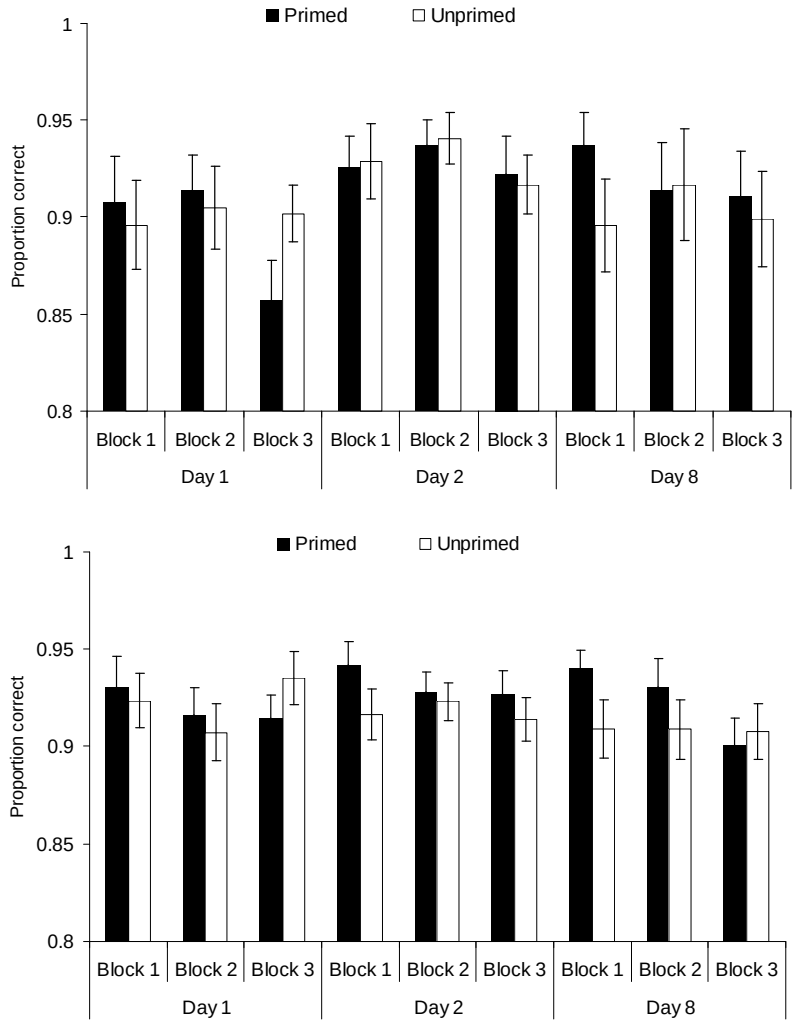


Figure 48. Accuracy rates in primed lexical decision with novel word primes, broken down by block and testing condition (upper panel = tested, lower panel = untested). Error bars represent standard error of the means.

In summary, averaged across the two testing groups and all blocks, a reliable priming effect in RTs was found on day 8 only. The same was the case with error rates. The untested group showed this same pattern in RTs, while priming in the in the tested group did not reach statistical significance. This is not surprising however as the group sizes were halved compared to the full initial analysis, resulting in reduced power. Visual inspection of Figure 46 suggests that there was a small numerical RT priming effect on day 1, and that it was carried by the tested group. Indeed, the by-block analysis showed a significant priming effect in the first block of day 1 in the tested group, both in RTs and accuracy rates. No effect on day 1 was found in any block in the untested group.

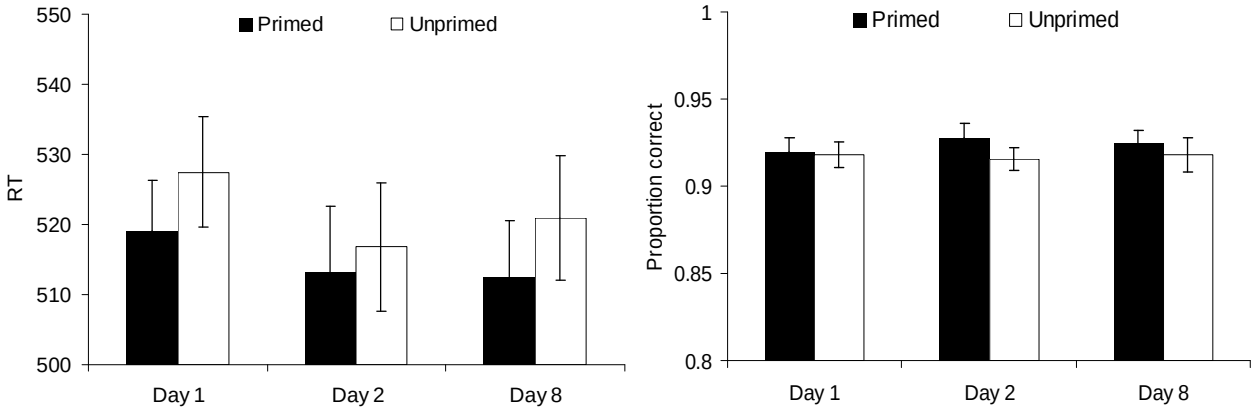


Figure 49. RTs and accuracy rates in the primed lexical decision task with real word primes and real word targets. Error bars represent standard error of the means.

Lexical decision with real word primes. Figure 49 (left panel) shows RTs to real word targets when preceded by a semantically associated (primed) or unassociated real word (unprimed). The data were trimmed in the same way as in the novel word prime condition. A mixed-effects linear model with subjects and items as random factors, and priming (primed vs. unprimed) and time of testing (day 1, day 2, day 8) as the fixed factors benefitted from subject-specific slopes for trial position, and item-specific slopes for counterbalancing list. The analysis showed no significant interaction between the two factors. The simplified model showed a significant effect of priming ($b = 0.014$, $t = 2.79$, $p = .02^{\dagger}$), with faster RTs to primed

trials. Time of testing also had a significant effect on RTs, with responses becoming faster from day 1 to day 2 ($b = -0.019$, $t = -5.39$, $p < .001$), from day 1 to day 8 ($b = -0.014$, $t = -4.00$, $p < .001$), but not from day 2 to day 8. The lack of interactions suggests that the priming effect is robust over time, but to get a more thorough picture of the effect, it was analysed for each day individually also. There was a significant difference between primed and unprimed trial on day 1 ($b = 0.016$, $t = 2.54$, $p = .01^{\dagger}$) and day 8 ($b = 0.015$, $t = 2.41$, $p = .02^{\dagger}$), but the effect failed to reach significance on day 2. The effect of time of testing in primed and unprimed conditions was looked at also. RTs became faster in the primed condition from day 1 to day 2 ($b = -0.016$, $t = -3.14$, $p = .002$) and from day 1 to day 8 ($b = -0.014$, $t = -2.76$, $p = .004$) but not from day 2 to day 8. The same was true of the unprimed trials (day 1 vs. day 2: $b = -0.022$, $t = -4.48$, $p < .001$, day 1 vs. day 8: $b = -0.015$, $t = -2.91$, $p = .003$). Hence this condition successfully demonstrated a priming effect which was largely unmodulated by time of testing, confirming that masked semantic priming can be observed in the current paradigm.

The accuracy rates are shown in the right panel of Figure 49. A mixed-effect logistic regression model with the same random and fixed factors (with subject-specific slopes for time of testing) showed no interaction between time of testing and

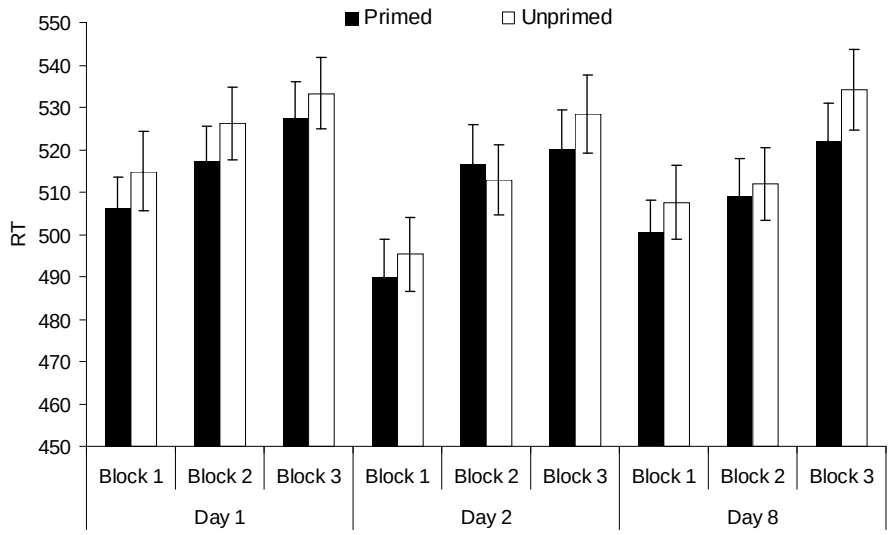


Figure 50. RTs in the primed lexical decision task with real word primes and real word targets, broken down by block. Error bars represent standard error of the means.

priming. The simplified model showed a priming effect that was marginally significant (b = -0.121, z = -1.89, p = .06†), and no effects of time of testing. When the priming effect was evaluated on each day separately, the effect did not reach significance on any day. The effect of time of testing showed no differences between day when looked at separately for primed and unprimed conditions.

As in the novel word prime condition, the real word prime condition was also analysed by dividing the data into three blocks to see if block modulated any of the effects (Figure 50). To do this, block (with three levels: block 1, block 2, block 3) was added as a fixed factor to the model described above. Subject-specific slopes for block were added. The model showed no significant three-way interactions, hence this term was dropped. Of the two-way interactions only that involving time of testing and block showed significant contrasts. This interaction showed that the difference between day 1 and day 2 RTs was larger in block 1 than in block 3 (b = -0.019, t = -2.21, p = .03†). Visual inspection of Figure 50 however suggested that RTs appear to grow slower in the later blocks. This effect of block only reached significance on day 2, where RTs increased from block 1 to block 2 (b = 0.030, t = 3.15, p = .002†), and from block 1 to block 3 (b = 0.042, t = 3.30, p = .001†). This effect was assessed also individually in the two priming conditions. In the primed trials, on day 1 RTs increased from block 1 to block 3 (b = 0.028, t = 2.02, p = .046†). On day 2 primed RTs increased from block 1 to block 2 (b = 0.035, t = 3.11, p = .003†) and from block 1 to block 3 (b = 0.039, t = 2.76, p = .006†) but not from block 2 to block 3. On day 8 the increases did not reach significance. In the unprimed condition there were no significant differences between blocks on day 1. On day 2 on the other hand unprimed RTs increased from block 1 to block 2 (b = 0.025, t = 2.22, p = .03†) and from block 1 to block 3 (b = 0.045, t = 3.22, p = .002†). On day 8 unprimed RTs increased from block 1 to block 3 (b = 0.030, t = 2.13, p = .04†) and from block 2 to block 3 (b = 0.032, t = 3.15, p = .002†). Finally, although priming did not interact with block, the priming effect was evaluated in each block and in each day separately to assess its strength. On day 1, the priming effect was significant in block 1 (b = 0.021, t = 2.22, p = .03†) but failed to reach significance in the other two blocks. On day 2 the effect was non-significant in all blocks. On day 8 the effect reached significance in the third block only (b = 0.026, t = 2.72, p = .007†).

Accuracy data were also examined as a function of block by adding block as a fixed factor in the logistic regression model described above. The data are shown in Figure 51. No three way-interactions reached significance, hence this term was dropped. No two-way interactions reached significance either, suggesting that block did not modulate the effects of priming or time of testing. When the priming effect was tested in each block on each day separately, a marginally significant priming effect was only found in block 2 of day 2 (b = -0.364, z = -1.88, p = .06†).

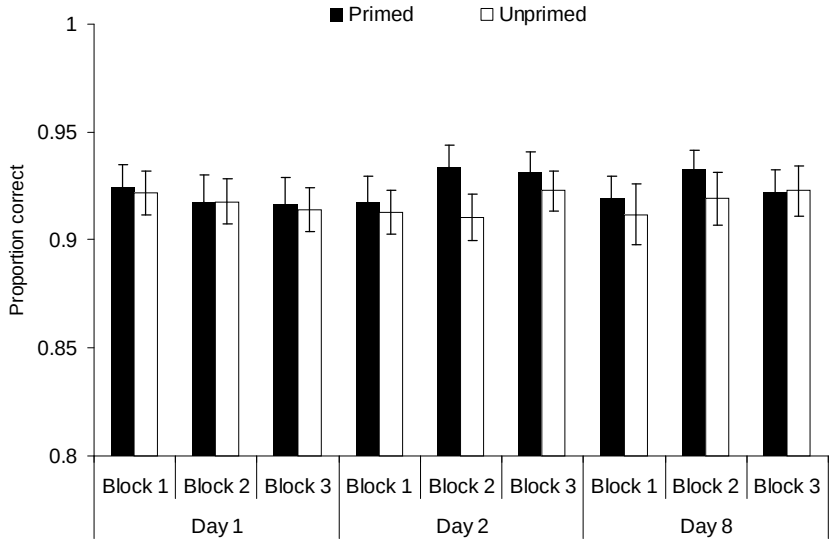


Figure 51. Accuracy rates in the primed lexical decision task with real word primes and real word targets, broken down by block. Error bars represent standard error of the means.

In summary, in the real word prime condition a masked priming RT effect was found without a significant interaction with time of testing, suggesting the effect was present in all testing sessions. A marginal effect was found in accuracy rates too. The by-blocks analysis did not reveal any surprising shifts in performance as the task progressed.

5.4.3 Discussion

The aim of Experiment 7 was to see if novel words could elicit priming in a task tapping into automatic semantic activation by using a masked prime with a short SOA. The real word prime condition confirmed that the task worked: a significant 7 ms priming effect was found when averaged across the three testing times. Priming

also had a marginally significant effect on lexical decision accuracy. The majority of participants were completely unaware of the existence of the prime, suggesting that the masking procedure was successful, although this conclusion relies on a subjective assessment, and cannot rule out partial awareness (Kouider & Dupoux, 2004). Nonetheless, it is reasonable to assume that this task involved a larger contribution of automatic processing than the task used in Experiment 6.

The time course with which the priming effect emerged in the novel word prime condition here was similar to the strategic priming task of Experiment 6 in that both effects required a delay between training and testing to emerge, although in the present experiment the effect was not yet seen after a short consolidation opportunity. Interaction contrasts showed that the priming effect was significantly larger on day 8 than on day 2. The contrast between day 8 and day 1 did not reach significance, however the priming effect was significant only on day 8 (8 ms), not on day 1 (5 ms) or day 2 (1 ms). This suggests that although there may have been weak priming present earlier, the effect grew stronger over the course of one week, with no significant development within the first 24 hours. This is consistent with Experiment 6, where the effect emerged only after offline consolidation had taken place. The accuracy data supported this late emerging RT effect with a small but significant priming effect on lexical decision accuracy found only on day 8. Although the priming effect on day 8 was numerically quite small, it is important to relate it to the 7 ms effect with real word primes: the novel word primes gave rise to a priming effect of the same magnitude as real word primes.

Comparison of the priming effect between the group that was tested on explicit meaning recall in the beginning of each test session and the group that was tested only at the very end of the experiment shed further light on the development of the priming effect. In the untested group there was no effect on day 1 or day 2, with a significant effect emerging on day 8. Unfortunately the priming effect failed to reach significance at each day in the tested group, however it is important to note that the numerical priming effect seen on day 1 appears to have been carried mainly by the tested group, who showed an 8 ms priming effect on day 1 and day 8. It is possible that the numerical priming effect on day 1 was caused by the tested group having had an opportunity to access the meanings just before carrying out the priming task. This may have given them an advantage in priming by allowing the novel meanings to have been recently activated in memory. This hypothesis was

supported by looking at the priming effect in each of the three blocks. Here the tested group showed a significant priming effect in the first block of day 1, with the effect failing to reach significance in the later blocks on that day. The untested group showed no priming effect in any of the blocks on that day. No such block-dependent priming effect was seen on the other days though, suggesting that perhaps the recall advantage can only be detected when combined with a preceding extensive training session.

The demonstration of a masked priming effect using novel word primes also provides important further evidence in favour of lexical integration. Lexical competition as a measure of lexical integration can be criticised due to the explicit overlap between the novel words and existing words. In a lexical decision task in particular it is possible that lexical competition effects emerge because participants are intentionally withholding their response on items that resemble the novel words. Such effects may not be a reflection of normal lexical competition. The masked priming effect on the other hand provides a task where lexical integration is seen even in a task where the participant is not consciously aware of the novel word, and no decision about the identity of the novel word is required. Hence the masked priming effect may be the strongest piece of evidence of lexical integration seen so far.

Another interesting set of data in the present experiment was provided by the explicit meaning recall task. In the other experiments reported in this and the previous chapters there has been significantly lower recall rates for novel words learned one or more days prior to testing, compared to words learned on the day of testing. As discussed earlier, this may have reflected either forgetting over time, or interference from the more recently learned set of novel words. In the current experiment participants learned only one set of novel words, eliminating the interference account. Here no change in meaning recall was seen from day 1 to day 2, suggesting that there is little or no decay over the course of one day when there is no interference from a new set of words. It should also be noted that once again no improvement in recall as a result of offline consolidation was seen, however as recall was near ceiling such an effect would be difficult to obtain. Recall rates did however decline significantly over a longer delay of one week from day 1 to day 8. Interestingly this decay appears to have been accompanied by the emergence of the priming effect, further enforcing the idea that the priming measured here was

independent of explicit recall and probably largely independent of strategic effects. The untested group seemed to have declined even more, as their recall performance on day 8 was significantly lower than that of the tested group. This implies that continued testing did slow down forgetting, but did not affect the emergence of priming.

5.5 Chapter Summary and General Discussion

The experiments reported in this chapter focused on the phenomenon of semantic priming, and on the question of whether newly learned words are integrated in the mental lexicon to a sufficient degree to allow semantic priming to occur. While a handful of earlier studies have addressed the same issue, the present experiments extend the earlier work in a number of ways. Firstly, three of the relevant earlier studies used semantic decision rather than primed lexical decision (Perfetti et al., 2005; Mestres-Misse et al., 2007; Mestres-Misse et al., 2008). Although the semantic decision task clearly requires access to word meanings, and measures the speed with which this access occurs, it is also likely to be influenced by a number of strategic processes. For example, it is possible that "primed" responses (responses confirming a relationship between the prime and the target) are faster than "unprimed" responses (responses indicating no relationship between prime and target) simply because making a positive response only requires the detection of the presence of a relationship, while a negative response requires a search of semantic memory before the response can be made with any degree of confidence. Seeing an RT difference between related and unrelated trials then may not reflect speed of access to meaning in general. The same concern can be raised against the N400 effect in this task, if the N400 is taken as a measure of the ease with which a word is integrated into preceding context. The primed lexical decision task used in the present experiments on the other hand is unaffected by this criticism as no explicit decision or response is required regarding the possible relationship between the novel word prime and a real word target.

Cross-modal priming studies by Breitenstein et al. (2007) and Dobel et al. (in press) avoided this problem by asking responses only to a picture target primed by an auditory novel word. However, here the prime-target pairs were also presented during training, with the meaning of the novel word acting as the target in the

priming test task. Hence it is difficult to say how much of what was being tested was episodic or semantic priming. A similar problem was present in the priming experiments of Dagenbach et al. (1990) where participants first learned the meanings of a set of novel words, and then learned episodic word pairs where the novel word was paired with a related real word.

The experiments presented in this chapter went beyond the existing literature by for the first time training participants on meaningful novel words and then seeing if these novel words can prime familiar words that were associated with the novel word meanings, rather than seeing if the novel words can prime their own definitions. Furthermore, as these experiments did not require participants to make any explicit response regarding the identity of the novel words, the present data for the first time test the activation of novel word semantics in exactly the same way in which semantic access is measured in familiar words in semantic priming studies. The key questions the present experiments attempted to answer were whether novel words are integrated in the lexicon such that they can prime familiar associated words through spreading semantic activation, and whether this is the case immediately after training or only after a period of offline consolidation.

Experiment 6 examined these questions in a priming task with a visible novel word prime and a long SOA. These two features of the task should have allowed participants to engage strategic processes alongside with automatic semantic priming. In this task semantic priming was only seen reliably in a condition where offline consolidation had been given time to occur between training and testing. There was no evidence that a long consolidation opportunity of one week provided an additional benefit compared to a short consolidation opportunity of about 24 hours. Interestingly though the novel word priming effect in this experiment (7 ms) was smaller than in the real word control condition (18 ms), suggesting that a completely normal priming effect may not have been reached within the time course of the experiment (although note that the difference is also likely to be affected by the different stimulus sets).

Experiment 7 used a masked prime and a short SOA to minimise the use of strategic processes, and maximise the contribution of automatic semantic activation. Remarkably, in this experiment a priming effect identical in magnitude to the real word priming effect emerged one week after training. No significant evidence of priming was seen immediately after training or one day after training, again

suggesting a gradually developing effect. This finding was consistent with the picture-word interference (PWI) data reported by Clay et al. (2007). The PWI task is similar to masked priming in that neither task requires a response or a decision to be made about the prime (the printed word in the PWI task). Furthermore, as the word in the PWI task is supposed to be ignored by the participants, any effect is has on naming performance is taken to be the result of automatic activation of the printed word distracter (although in the PWI task participants remain aware of the words, whereas in the masked priming task they are largely unaware of the prime). However, Clay et al. (2007) tested their participants immediately after training and one week after training, with no intervening tests. This means the effect may have emerged earlier, and may in fact benefit from sleep, like the other language learning tasks discussed in Chapter 3 (e.g., Dumay & Gaskell, 2007; Gomez et al., 2006; Fenn et al., 2003). However, by looking at priming both after 24 hours and one week, Experiment 7 suggested that automatic priming continues to develop beyond one day or night of consolidation.

Another finding that highlights the difference between Experiments 6 and 7 was the influence of block on priming. In Experiment 6 the strongest priming effect in consolidated novel words was found on the third block of the task (although the interaction with block did not reach statistical significance). In Experiment 4 a consolidation advantage in semantic decision only emerged in the third block of the task. It appears then that tasks that involve explicit access to novel word meanings benefit from repetition over the course of the task. The priming task in Experiment 7 on the other hand did not involve explicit access to novel word meanings, and consequently no evidence was seen of the later blocks enhancing the priming effect.

While both priming experiments showed that offline consolidation plays a significant role in the development of semantic priming in novel words, the time courses in the two experiments were slightly different, in that Experiment 6 showed a reliable effect already one day after training, while in Experiment 7 the effect reached significance one week after training. This type of difference between the time it takes for strategic and automatic priming effects to emerge may reflect different methods or routes of accessing word meanings. For example, the multistage activation model of Stolz and Besner (1996) proposes a lexical level from which activation spreads to a semantic level. In a model such as this the formation of a new lexical representation and the linking of that representation to the semantic level is

the process that is likely to be affected by consolidation. Automatic semantic activation would require a fully functional and reliable connection between the new lexical representation and the semantic level. Strategic or explicit retrieval of the novel word meaning on the other hand might be accomplished even via a connection that is not yet fully consolidated. Distributed network models (e.g., Plaut & Booth, 2000) can also conceivably accommodate a consolidation process. A weakly activated semantic representation would not be able to activate an overlapping similar representation in the absence of some sort of top-down boost from explicit recall. However, as consolidation progresses by the reinstatement mechanism postulated by the CLS accounts, the novel representation gains in strength and will eventually activate overlapping representations in an automatic fashion without extra input from explicit recall. Both of these suggestions imply that for automatic activation to occur, a more profound change may need to take place in semantic memory. Experiment 7 and the data from Clay et al. (2007) suggest that such a change takes several days to happen while a more strategic alternative route for activating the new meaning may become available earlier, and may or may not benefit from sleep in particular. This is not to say that the consolidation that operates on automatic access does not benefit from sleep but it does imply that it would require several nights of sleep.

Experiment 6 suggested that another task that may benefit from more than one day or night of consolidation is shadowing. In Experiment 4 a consolidation benefit was seen for words that had been learned one day earlier suggesting that even one day of consolidation can be important in this task. The effect was however not seen in Experiment 5 again with a one day consolidation opportunity. In Experiment 6 on the other hand a consolidation benefit was seen only in participants who had had a long consolidation opportunity of one week. It seems then that this task involves a consolidation process that operates over several days, and that sometimes the effect can already be seen after one day of consolidation, but on other occasions more than one day is needed. There are at least two potential reasons why this task might benefit from long term consolidation. One reason is that shadowing may involve a strong semantic component. Meaning does seem to influence shadowing, as mentioned in Chapter 4 Slowiaczek (1994) showed semantic priming effects in shadowing. If this is the case, then the consolidation benefit seen in Experiment 6 may have been due to semantic support gradually becoming available over the

course of the week. The second reason might be that shadowing in Experiments 4 and 6 required access to phonological representation of the novel words, but the phonological forms of the novel words had never been heard before by these participants. In other words, the phonological representation being accessed in this task had been generated as a by-product of purely orthographic exposure. This may result in a weak phonological representation that requires a period of offline consolidation to gain in strength. This may be a process that continues over several days or nights, possibly because the reinstatement process is slowed by the absence of direct experience with the phonological word forms. These hypotheses remain tentative at this point, and would require direct empirical investigations to develop further.

Finally, Experiments 6 and 7 together shed more light on the influences of time on explicit recall of novel word meanings. Experiments 4-6 demonstrated a pattern where explicit recall seemed to suffer from passing time. Recall was always better for words learned immediately before testing compared to words learned a day before testing. As discussed earlier, this may have reflected either forgetting over time, or forgetting due to interference from learning a second set of novel words. In Experiment 7 participants learned only one set of novel words, whose recall was tested immediately after training, one day after training, and one week after training. Under these circumstances no significant forgetting was seen within one day. This seems to suggest that the forgetting seen in the earlier studies was mostly due to interference. Significant forgetting was seen one week after training, although even then the forgetting was much smaller than in the earlier experiments. In Experiment 7 performance declined from 96% of objects recalled after training to 94% recalled one day later, a decline of only 2%. Contrast this with the difference of 27% in Experiment 6 between words learned the day before and words learned on the day of testing in the short consolidation group. It must be noted though that the small decline in Experiment 7 is also likely to be partially due to repeated testing on day 1 and day 2, whereas in Experiment 6 testing took place only once.

To summarise, the main finding of Chapter 5 was that novel word meanings do appear to benefit from offline consolidation, but the consolidation period is longer than seen in tasks measuring word form knowledge, where the first night of sleep may be the key (Dumay & Gaskell, 2007). Specifically, activation that relies predominantly (but probably not exclusively) on strategic access benefits from

consolidation over the first 24 hours following training, but seems to benefit less from further consolidation beyond that. Automatic activation on the other hand seems to benefit from consolidation continuing up to a week, at least when measured by semantically primed lexical decision. Taken together, these experiments show that unlike word form learning, learning word meanings engages a gradual consolidation process that continues for more than one day or night of sleep. This is in contrast with the ERP studies of Perfetti et al. (2005) and Mestres-Misse et al. (2007) which found an N400 effect with novel words immediately after training. This may suggest that the N400 is more sensitive to episodic memory traces than the priming paradigms used here, or that the methodological issues discussed in Chapter 3 can have a significant impact on the N400. Shadowing, a task that may well also involve a significant semantic component, also appears to be sensitive to consolidation over several days. The design used in the current experiments did not allow any conclusions to be drawn about the role of sleep specifically on consolidation of semantic information in novel word learning. This question is left for future studies to resolve. In the next chapter however I will address the role of sleep in consolidation of novel word forms in the absence of given meaning, and the specific neural events that may participate in sleep-associated memory consolidation.

Chapter 6: Lexical integration and the architecture of sleep

6.1 Introduction

As reviewed in Chapter 3, there is now plenty of evidence to suggest that newly learned linguistic materials benefit from offline consolidation, and that sleep appears to play a crucial role in this consolidation process. While the experiments reported in the previous chapter suggest that learning the meaning of novel words triggers a consolidation process that operates over several days and/or nights, the acquisition of novel word forms seems to depend most on the first night of consolidation. Evidence for this was seen in some of the experiments reported earlier in this thesis. Experiment 5 for example showed that participants in a cued recall task recalled more novel words that had been learned on the previous day compared to words learned just before testing. Sleep was not manipulated in that experiment, however presumably all participants slept during the 24 hours between training and testing.

As already mentioned in Chapters 1 and 3, Dumay and Gaskell (2007) attempted to tease apart effects of consolidation occurring during wake and sleep. Participants who were trained in the evening and tested in the morning after a night of sleep showed lexical competition effects and improved free recall performance. Participants who were trained in the morning and tested in the evening, with presumably no intervening sleep, showed no lexical competition effect and no improvement in free recall. While these data suggest that sleep plays a key role in integrating the novel words in the existing lexicon, they also raise a question about the environmental and neural circumstances under which consolidation takes place: is sleep beneficial in consolidation because it provides an environment free of external interference, or are there neural events that take place during sleep that drive consolidation?

One way to address these questions is to look at sleep architecture, and try to see if there are some physiological aspects of sleep that are correlated with consolidation. The most common approach has been to look at global sleep architecture, that is, the involvement of different sleep stages in consolidation. Another emerging target of research in sleep and memory consolidation is sleep

spindles. I will give a brief overview of the literature on these two approaches in the next two sections.

6.1.1 Memory consolidation and sleep stages

Sleep can be divided into different sleep stages, most importantly rapid eye movement (REM) sleep, and non-REM (NREM) sleep which comprises of slow wave sleep (SWS: sleep stages 3 and 4), and lighter sleep stages 1 and 2. The sleep stages appear to have different roles to play in consolidating different types of memories. According to the dual process theory (see e.g., Diekelmann, Wilhelm, and Born, 2009, for a review), SWS supports consolidation of declarative memories, while REM supports consolidation of procedural memories. Although early work using REM deprivation supported this notion, sleep deprivation as a methodology is prone to confounds (e.g., due to the general cognitive impairment resulting from sleep deprivation), and hence the most often cited evidence for the dual process theory comes from the work of Plihal and Born (1997, 1999) who used the split-night paradigm. This approach takes advantage of the fact that sleep during the first half of the night is dominated by SWS, and sleep in the second half of the night is dominated by REM. For example, Plihal and Born (1997) trained participants on a word-pair list (declarative task) either in the beginning of the night or in the second half of the night, followed by three hours of either SWS rich early sleep or REM rich late sleep, after which they were tested. While both groups showed a sleep-associated increase in recall, the SWS group improved significantly more than the REM group. The opposite pattern was seen when a mirror tracing task (procedural task) was used. However, further research has revealed a more complicated picture. Gais, Plihal, Wagner, and Born (2000) showed in a split-night paradigm that SWS-dominated early sleep improved visual texture discrimination, a form of procedural knowledge. Declarative materials have also been shown to benefit from REM sleep, especially if they are highly emotional (Wagner, Gais, & Born, 2001). There is also evidence that stage 2 sleep is important, at least in the procedural domain. Walker, Brakefield, Morgan, Hobson, and Stickgold (2002) showed that overnight improvement in a sequential finger tapping task correlated with time spent in stage 2 sleep, particularly during the latter half of the night. This variability in the data has led many researchers to call for a move away from focusing on sleep stages, and for

an examination of the physiological processes that occur during sleep, and different types of learning in more detail (Gais & Born, 2004; Fogel, Smith, & Cote, 2007).

6.1.2 Memory consolidation and sleep spindles

In recent years there has been much interest in the role sleep spindles play in memory consolidation. Sleep spindles are bursts of fast oscillations (~11-15 Hz) that last at least 0.5 s but not usually more than 3 s, and occur during NREM sleep. The most intriguing aspect of spindles is that they have been demonstrated to occur in a temporally synchronised manner with hippocampal ripples (brief high frequency bursts of activity in the 100-200 Hz range), both in animals (Siapas & Wilson, 1998; Sirota, Csicsvari, Buhl, & Buzsaki, 2003) and in humans (Clemens, Molle, Eross, Barsi, Halasz, & Born, 2007), and hence may be involved in the hippocampal-neocortical transfer of newly acquired memories postulated by the complementary learning systems (CLS) accounts.

As reviewed in Chapter 3, CLS accounts argue that new memories are initially dependent on the fast learning hippocampus, which supports the new memory trace as it becomes consolidated in the slow learning neocortex. During offline periods, such as sleep, the neocortical memory trace is reinstated repeatedly until it becomes independent of the hippocampus, allowing the hippocampal representation to decay. As hippocampal ripples constitute the most prominent neural event in the hippocampus during sleep, it has been hypothesised that these spikes of activity play a key role in the reinstatement process. Compelling evidence of this was recently provided by Girardeau, Benchenane, Wiener, Buzsaki, and Zugaro (2009) in a study where hippocampal ripples were blocked in rats learning to find food in a radial maze. Rats whose hippocampal ripples were suppressed during sleep following learning trials learned slower and failed to reach the same level of performance as control rats in which the suppression did not target ripples. Sirota et al. (2003) showed in rats that hippocampal ripples and cortical sleep spindles are closely temporally coupled. These authors suggested that spindles select the hippocampal cells that will participate in the ripple event, which in turn provide output to those cell assemblies which participate in the spindle.

Spindles' involvement in memory consolidation is supported by a large amount of behavioural data. In the procedural domain several authors have showed

an association between spindle activity and degree of performance enhancement over a sleep period, looking at skills such as spatial navigation (Meier-Koll, Bussmann, Schmidt, & Neuschwander, 1999), visuospatial memory (Clemens, Fabo, & Halasz, 2006), and motor learning (Milner, Fogel, & Cote, 2006; Fogel & Smith, 2006; Fogel, Smith, & Cote, 2007; Nishida & Walker, 2007; Morin et al., 2008; Tamaki, Matsuoka, Nittono, & Hori, 2008, 2009). Similar associations have been seen in declarative tasks as well, including word-pair learning (Gais, Molle, Helms, & Born, 2002; Schabus et al., 2004; Schmidt et al., 2006; Schabus et al., 2008) and face-name association (Clemens, Fabo, & Halasz, 2005).

In addition to the correlational evidence showing an association between learning and subsequent spindle activity, there have also been demonstrations of the nature of the learning materials affecting spindle activity in specific ways, showing that spindle activity is sensitive to the materials to be consolidated. Schmidt et al. (2006) showed that compared to a non-learning task spindle activity increased only after participants had learned abstract word-pairs, while no effect on spindles was seen as a consequence of learning concrete word-pairs, showing that task difficulty modulates spindle activity. Spindle activity also appears to be regionally specific to the area of the cortex which is most involved with processing the materials to be learned. Nishida and Walker (2007) trained participants on a finger tapping task using the left hand. Spindle activity during the following nap period was correlated with the amount of performance improvement, but only when measured at electrodes over the contralateral (right) motor area. This is further evidence that spindles are involved with consolidating specifically the recently learned experiences.

Finally, it is worth noting that while the early spindle studies examined sleep spindles in the ~11-15 Hz range, there is now emerging evidence for two different spindle types: slow spindles (~11-13 Hz) and fast spindles (~13-15 Hz). Not only are these two types typically observed in different areas of the cortex, with slow spindles dominating frontal areas and fast spindles dominating parietal areas (Schabus et al., 2007), but they also appear to be important in consolidating different types of material. In the declarative face-name association task Clemens et al. (2005) found spindle correlations in the left frontal electrodes, while in a visuospatial task the same research group found spindle correlations in the parietal electrodes (Clemens et al., 2006). These studies did not distinguish between spindles in the different frequency ranges, but it is likely that the frontal spindles corresponded to slow

spindles, and the posterior activity corresponded to fast spindles. The suggestion that declarative learning is associated with frontal slow spindles was further supported by the study mentioned above by Schmidt et al. (2006), where the spindle correlations were found only for slow spindles, and on frontocentral electrodes. In contrast to the association between slow/frontal spindles and declarative learning, procedural learning seems to be most reliably associated with fast/parietal spindles. Tamaki et al. (2008, 2009) found that learning a visuospatial motor task increases fast spindle activity compared to a non-learning condition, but not slow spindle activity. Finally, Milner et al. (2006) found increased spectral power in only the fast spindle range after participants learned a motor task.

In sum, both sleep stage and sleep spindle data provide promising avenues of research for looking at the neural events that drive memory consolidation during sleep. Applied to word learning tasks, it can be hypothesised that different sleep stages may be involved in consolidating different aspects of word knowledge. Based on the demonstration of a link between SWS and declarative learning (such as learning word-pairs), tasks measuring explicit recall or recognition of novel word forms or meanings should benefit from SWS. The declarative vs. procedural dichotomy is less helpful in making predictions about emergence of lexical competition effects though. In the lexical competition paradigm participants are not required to make decisions or recognise the newly learned words, instead what is measured is the indirect influence of the new words on the recognition of phonologically overlapping familiar words. Such an effect which involves both explicit and implicit components may be associated with either SWS or REM sleep, or both. In the sleep spindle literature there are no studies looking specifically at the integration of new memories with existing memories, which is what the lexical competition paradigm measures. However, the CLS accounts suggest that integration of new memories is one of the most crucial results of offline consolidation and hippocampal-neocortical transfer, hence it is reasonable to expect spindle activity to be closely associated with emerging lexical competition effects. The declarative tasks that have been shown to be associated with spindle activity also suggest that spindles may facilitate consolidation of explicit word recall as well, but perhaps not as strongly as lexical integration.

6.2 Experiment 8

The main purpose of Experiment 8 was to see which aspects of sleep architecture (specifically sleep stages and spindles) are associated with memory consolidation in learning of novel word forms, looking at both lexical competition and explicit recall and recognition measures. Participants were trained on 30 spoken meaningless novel words, followed by tests of lexical competition and explicit recall and recognition immediately after training, about 10 hours after training, and again one week after training. Importantly, half of the participants were trained in the evening, and spent the night between the immediate and the delayed test in the laboratory, while polysomnographic measures were collected during sleep (sleep group). The other group were trained in the morning and tested in the evening, with no intervening sleep (wake group). Both groups were tested again one week later, at the same circadian time as the delayed test.

While Dumay and Gaskell (2007) also took measures of lexical competition in an experiment using a similar design, the current experiment made some important changes. Firstly, the current experiment used lexical decision to base words, while Dumay and Gaskell used pause detection. Both measures have been shown to reliably reveal lexical competition effects. Secondly, Dumay and Gaskell used novel words where the novel item was formed by adding a syllable to the end of a familiar word (e.g., *shadowks*, from *shadow*). The current experiment used novel words derived from existing words by changing the phonemes at the end of the words, starting at the final vowel (e.g., *cathedruke*, from *cathedral*). Again, these novel words have been shown in the past to give rise to lexical competition effects (e.g., Gaskell & Dumay, 2003). Thirdly, while Dumay and Gaskell tested participants immediately after training, 12 hours later, and again 24 hours later, the current experiment delayed the third test until one week after training, to see if any potential effects involving sleep architecture on the first night would still be seen one week later. The most important change to Dumay and Gaskell however was the collection of polysomnographic measures overnight.

Apart from changes to the design and materials used by Dumay and Gaskell (2007), Experiment 8 also introduced some new test tasks. In addition to lexical decision as a measure of lexical competition, this experiment used free recall, cued recall, and old/new categorisation tasks. Free recall was used also by Dumay and

Gaskell, and the expectation here was to replicate their finding of improving recall overnight in the sleep group, but no significant change in performance in the wake group. The cued recall task was added for two reasons. Recall that in Experiment 5 participants had higher cued recall accuracy rates in words learned on the previous day compared to words learned on the day of testing. This consolidation effect may have been the result of sleep-dependent consolidation, however as there was no sleep vs. wake contrast, it was not possible to determine whether sleep was of importance in that task. If sleep provides the optimal conditions for consolidation, there should be a benefit for the sleep group over the wake group in the delayed test. It should be noted however that by necessity the modality was changed in Experiment 8, and hence the composition of the cues was also different. The old/new categorisation task was designed to be an improved version of the two-alternative forced choice (2AFC) task used by Dumay and Gaskell. The 2AFC test results in very high levels of accuracy, possibly masking differences between the groups. Dumay and Gaskell found no effect in this test, although Davis et al. (2009) did find a difference between consolidated and unconsolidated novel words. However, the old/new categorisation task is possibly more sensitive to consolidation not only because it is more difficult, but also because it allows an analysis of reaction times as well as accuracy rates. If this task benefits from sleep, it was expected that the sleep group's response times would become significantly faster overnight, while the wake group should show little improvement. The same should occur with accuracy rates.

Finally, it was expected that the same pattern of emerging lexical competition effects should be seen as in Dumay and Gaskell (2007), with no competition effect in either group immediately after training, and an effect seen in only the sleep group in the delayed test. Both groups should show the effect in the one-week follow up, at which point both groups would have been able to sleep prior to testing. As mentioned above, the most critical aspect of the data relate to the PSG measures, with potential correlations emerging between measures of word learning and sleep stages or sleep spindle activity.

6.2.1 Method

Materials

The critical stimuli consisted of 60 base words, chosen from a set of 68 words used by Tamminen and Gaskell (2008). The stimulus selection here was done on the basis of the likelihood of observing the lexical competition effect: the items that were used were the ones that resulted in the largest lexical competition effect in Tamminen and Gaskell (2008). All chosen words were bisyllabic (n = 31) or trisyllabic (n = 29), with a phoneme length of 8.0 on average (range = 6-11). The mean frequency was 4.5 occurrences per million (range = 2-18). All base words had an early uniqueness point, located before the final vowel.

Each base word had two corresponding novel words which diverged from the base word at the final vowel (e.g., *cathedruke* and *cathedruce* derived from *cathedral*). One was used as the trained novel word, and the other as a foil in the old/new categorisation task. Note that these two novel words differed only by one phoneme, which was always the final one. This was done to make the task more challenging. All novel words and foils were taken from Tamminen and Gaskell (2008). All base words and novel word stimuli are presented in Appendix 10.

Sixty real words were selected to act as fillers in the lexical decision task. This meant that experimental base words (base words for which a new competitor was trained) made up only 25% of the real words (30 experimental base words, 30 control base words for which no new competitor was trained, 60 filler words). The filler words were all monomorphemic nouns, and included monosyllabic (n = 30), bisyllabic (n = 15), and trisyllabic words (n = 15). All filler words had frequencies less than 50 occurrences per million (M = 6.4, range = 2-20), and were of similar length to the base words (M = 5.2, range = 4-9).

Finally, 90 nonwords were created for the lexical decision task. These consisted of 30 monosyllabic, 30 bisyllabic, and 30 trisyllabic nonwords. All nonwords were created by taking a real word (not used in the experiment) and changing one phoneme which could be either in any position. They were similar in length to the real words (M = 5.6, range = 4-9). Ten items, five words and five nonwords, were also selected for a practice block. All fillers were selected from the pool of filler items used by Tamminen and Gaskell (2008).

As this experiment was carried out in Boston, USA, all stimuli were recorded by a female native speaker of North American English, in a sound proof booth using the same recording equipment as in the previous experiments. In addition, a different female speaker of North American English recorded the novel words and corresponding foils to be used in the old/new categorisation task. The reason for using a different speaker in this task was to try to stop participants relying on episodic memory of the training stimuli when doing the categorisation task. For example, participants might rely on physical cues, such as speech rate or recording quality, when deciding whether *cathedruke* or *cathedruce* was a word heard in training. By using a different speaker the decision would have to be made based on the abstract lexical representation.

Three measures of subjective alertness were used to assess how sleepy participants felt at the time of carrying out the test tasks. The first two of these consisted of 115 mm visual analogue scales asking participants to rate their ability to concentrate, and how refreshed they felt at the time, by marking two lines printed on the questionnaire (with end points labelled "poor" to "excellent" for ability to concentrate, and "not at all refreshed" and "very refreshed" for how refreshed they felt). The third measure was the Stanford Sleepiness Scale (SSS; Hoddes, Zarcone, Smythe, Philips, & Dement, 1973), where participants were asked to rate their level of sleepiness on a scale from 1 to 7, which each point in the scale labelled as follows: 1 = Feeling active, vital, alert, or wide awake, 2 = Functioning at high levels, but not at peak, able to concentrate, 3 = Awake, but relaxed, responsive but not fully alert, 4 = Somewhat foggy, let down, 5 = Foggy, losing interest in remaining awake, slowed down, 6 = Sleepy, woozy, fighting sleep, prefer to lie down, 7 = No longer fighting sleep, sleep onset soon, having dream-like thoughts.

Design

The base words were divided into two sublists of 30 items in each: one list was used as the experimental list, i.e. base words for which a new competitor would be taught. The other list acted as a control list, i.e. base words for which no new competitor was taught. This allowed a comparison of the recognition times to words in the two conditions: words in the experimental list should have slower recognition times due to the newly learned competitor. Across all participants both lists were used in both conditions an equal number of time. It was important that the items in

the lists were matched in recognition times to minimise random statistical noise that might obscure the lexical competition effect. Hence recognition times to each item were taken from Tamminen and Gaskell (2008) and used as basis for checking that the lists were indeed matched on this variable.

The use of the old/new categorisation task presented a problem, as this task required exposing participants to the novel words after training, and hence may act as further training, affecting levels of performance in the other recall tasks in the delayed test and the one-week follow up test. Although this would be true for all participants and thus not confound the sleep vs. wake manipulation, it was decided to assess the problem directly by only exposing participants to half of the novel words in the old/new categorisation of the immediate test, and to the full set of novel words in the two subsequent tests. The two lists of 30 base words and their corresponding novel words were hence further divided pseudorandomly into two lists of 15 stimuli, with only one of these lists used in the old/new categorisation task immediately after training.

Procedure

Each participant was randomly assigned either to a sleep group or a wake group. Participants in the sleep group arrived in the laboratory at 19.30 on the first day of the experiment. They then filled in a consent form, a sleep log covering the last three nights, the Epworth Sleepiness Scale, and a general demographic form. The electrodes for polysomnographic recording were attached prior to the beginning of the training session. After the training session, which was initiated at about 21.00 and lasted about 60 minutes, but prior to starting the testing session, participants filled in the alertness and sleepiness questionnaires. The timing of the presentation of the questionnaires was chosen to allow me to take these measures at the time of testing rather than at the time of training, and to allow a comparison across evening and morning testing sessions. After completing the testing session (initiated on average at 22.15) which lasted about 30 minutes, participants slept overnight in a laboratory bedroom. Participants were woken up by the experimenter in the morning at 07.45, allowing for a sleep opportunity of about 8.5 hours. Participants were allowed to get dressed, have a small breakfast, and the electrodes were removed. The second testing session was initiated about 35 minutes after waking up, again preceded by filling in of the alertness questionnaires. A follow-up testing session,

identical in procedure to the other test sessions, was held on average 7 days later (range = 3-11 days) at approximately the same circadian time as the second testing session, on average at 10.15 (no earlier than 09.00 and no later than 11.00).

The procedure for the wake group was identical to that of the sleep group, except for the timing of the sessions. Participants in the wake group arrived in the laboratory at 09.00 on the first day of the experiment. They filled in the same questionnaires, sleep log, and consent form as the sleep group. They then completed the training session, followed by the first testing session (initiated on average at 09.50). Participants were then free to spend the day as they would normally, although they were asked to refrain from caffeine and alcohol during the day. They returned to the laboratory in the evening, and carried out the second testing session (initiated on average at 19.30). Note that the timing of the sessions was designed to match the intervening time between the first two sessions for both groups of participants as closely as possible within practical constraints. The mean time that elapsed between initiating the first test and the second test was 10 hours and five minutes in the sleep group, and 9 hours and 40 minutes in the wake group. The follow-up session for the wake group took place on average 7 days later (range = 3-7 days).

Training tasks. The training consisted of two tasks: phoneme monitoring and word repetition. In phoneme monitoring the task was to decide whether an auditorily presented novel word contained a predetermined target sound. In word repetition participants were simply asked to repeat aloud an auditorily presented novel word. While phoneme monitoring is a task used in most previous novel word studies looking at emergence of lexical competition, word repetition was added in this experiment in order to give participants a chance to get accustomed to saying the novel words aloud since two of the test tasks also required them to do so.

Each participant completed five main blocks of phoneme monitoring, each block consisting of six sub-blocks. In each sub-block only one of the six target phonemes was used (/p/, /d/, /s/, /m/, /n/, /l/), and each sub-block included one presentation of each novel word. Hence each novel word was heard a total of 30 times during the phoneme monitoring task. Each of the five main blocks of phoneme monitoring was interleaved by one block of word repetition resulting in total number of four blocks of word repetition with one presentation of each novel word per block. Hence by the end of the training session, each novel word had been heard 34 times.

The order of tasks and blocks was fixed, but the order of presentation of novel words within a block was randomised by the software used for stimulus delivery (E-prime).

A phoneme monitoring trial started with the visual presentation of the target phoneme on screen for 500 ms, followed by auditory presentation of the novel word via headphones. A response was made by pressing a key on a standard laptop keyboard, labelled "Yes" or "No". Participants were encouraged to respond quickly and accurately, with a response deadline of 3000 ms. A word repetition trial started with a visual warning message "READY" for 500 ms, followed by auditory presentation of the novel word. Participants were asked to repeat the novel word aloud and to press the Enter key to move on to the next trial. Responses to phoneme monitoring trials were recorded, but responses to the word repetition trials were not, although the participants were unaware of the latter. Stimulus presentation and response collection in the training and testing sessions was carried out on Dell laptops running Windows XP and E-prime. Auditory stimuli were delivered via Beyerdynamic DT 234 Pro headphones with an integrated microphone for recording vocal responses in the testing tasks. The laptop's keyboard was used as the manual input response device in all tasks.

Testing tasks. Testing sessions included a test of lexical competition (lexical decision), free recall, cued recall, and old/new categorisation. The order of the tasks was fixed. In the lexical competition task participants were asked to make a lexical decision to the experimental base words, control base words, filler words, and filler nonwords. A lexical decision trial started with the presentation of a fixation cross on the screen for 500 ms. This was followed by auditory presentation of the stimulus, which also started timing. Once a response was made, accuracy feedback was provided on screen in the form of a happy or a sad cartoon face. If no response was detected within 2500 ms from the offset of the word, a message saying "no response" was displayed. Feedback was displayed for 750 ms, after which a new trial was initiated. The order of presentation was randomised by E-prime. Participants were encouraged to always respond as quickly and as accurately as possible. Participants were not informed of the relationship between the base words and the novel words they had learned earlier.

In the free recall task participants were given 3 minutes to recall as many novel words as possible, without cueing or prompting, and to say them aloud. This part of the experiment was timed by the experimenter who remained in the testing

room for the duration of the task. Responses were recorded on a minidisc using the integrated microphone in the headphones. The responses were transcribed and scored offline by the experimenter.

In the cued recall task participants were presented with the first two or three phonemes of a novel word and asked to recall and say the complete novel word aloud. The cues were recorded by the same speaker who made the recordings used in the training tasks and were selected so that each cue could only correspond to one novel word. Each cued recall trial started with the presentation of a fixation cross on the screen for 500 ms, followed by the presentation of the auditory cue. Presentation order was randomised by E-prime. Participants were given 10 s to make a vocal response, and asked to make a key press after saying the word to move on to the next trial. If a new trial was not initiated within 10 s, a message was displayed on the screen asking the participant to say the word "blank" and to move on to the next trial. The vocal responses were again recorded on a minidisc, and transcribed and scored later by the experimenter.

In the old/new categorisation task participants were presented with novel words and the novel word foils in a pseudorandomly ordered list, and asked to decide after each stimulus whether it was a trained novel word (old), or a foil not heard in training (new). A trial started with the presentation of a fixation cross for 500 ms. The auditory stimulus was then presented, after which 3000 ms was given to make a response by pressing a key labelled "Yes" or "No" on the keyboard. Participants were instructed to respond Yes to trained novel words and No to foils. No accuracy feedback was provided in order to prevent this task from acting as a further training opportunity, but RT feedback was given in the form of a message saying "too slow" if no response was detected within the 3000 ms. Two pseudorandom orders of presentation were created with Mix (van Casteren & Davis, 2006) for each of the three testing sessions. In both orders half of the novel words were preceded by their corresponding foils, and vice versa. Furthermore, each novel word and its foil were separated by a minimum of four items, and no more than five trials of one response type (trained or foil word) were allowed to be presented in sequence. Different orders were used in each testing session, so that no participant experienced the same order more than once. Participants were instructed to respond as quickly and as accurately as possible.

Polysomnographic recording. A Grass Technologies system was used to record EEG at a 200 Hz sampling rate. Four scalp electrodes were used, positioned according to the international 10-20 system (F3, F4, C3, C4), with each electrode referenced to the contralateral mastoid. Two electro-oculographic (EOG) channels were used to monitor eye movements, and two electromyographic (EMG) channels monitored chin movements. Sleep data were categorised into sleep stages visually in 30 s epochs according to Rechtschaffen and Kales (1968). Table 4 shows the main sleep parameters of the participants in the sleep group. Participants slept on average 8 hours, with a mean sleep onset latency of 15 minutes. Average times spent in the different sleep stages in this experiment were comparable with values reported in other sleep studies.

Table 4. Sleep parameters in overnight participants in Experiment 8.

Sleep parameter	Mean time (min) ± SEM	% of total sleep time ± SEM
Total sleep time	478 ± 6	
Wake after sleep onset	17 ± 2	
Sleep latency	15 ± 3	
Stage 1	28 ± 2	5.8 ± 0.5
Stage 2	274 ± 6	57.5 ± 1.0
SWS (Stages 3 + 4)	77 ± 4	16.3 ± 0.9
REM	97 ± 5	20.3 ± 0.9

Note: SWS = slow wave sleep, REM = rapid eye movement sleep, SEM = standard error of the mean.

Participants

Sixty-five native English speaking participants were recruited for this experiment. All participants were required to abstain from alcohol and drugs for 24 hours prior to the experiment, and to avoid consuming caffeine during the day of training and between the first and second tests. Further exclusion criteria included medication affecting sleep, history of sleep disorders, and history of serious mental disorders. Participants were asked to maintain a regular sleep schedule on the three days prior to the experiment. This was confirmed by asking participants to fill in sleep logs covering the past three nights upon arriving in the laboratory. One participant dropped out after the first session, and two further participants dropped out after two sessions. Data from the former has been excluded, but the latter datasets have been retained as overnight consolidation can still be evaluated for these participants. One participant's data in the sleep group were excluded due to non-compliance with the recruitment criteria, and another in the wake group was

excluded due to chance level performance in the training task (52% correct in phoneme monitoring), suggesting this participant did not comply with the training instructions. This left 62 participants in total, all of whom were students at colleges in the Boston (USA) area. Thirty-one participants served in the sleep condition (8 males, mean age = 20.4, range = 18-30), and another 31 in the wake condition (10 males, mean age = 20.5, range = 18-30). No participants reported language disorders, or had participated in any of the previous experiments reported in this thesis. Participants were paid $50 (about £30) for taking part.

6.2.2 Results

6.2.2.1 Behavioural data

Training. Accuracy in the phoneme monitoring task was analysed to make sure all participants had attended to the task, and to see if the sleep or wake group showed evidence of better learning during the training. The sleep group had an average accuracy rate of 84.5% (SEM = 0.02), while the wake group scored 86.3% (SEM = 0.02) correct. A mixed-effects logistic regression model was fitted to the data. Subjects and items were included in random effects, and group (sleep vs. wake) as a fixed factor. LLR tests showed that subject-specific random slopes for trial position improved goodness of fit. The effect of group on accuracy failed to show a significant effect ($b = 0.042$, $z = 0.23$, $p = 0.81$). It appeared then that the participant groups were equally successful in the phoneme monitoring task.

Testing. Data from the lexical decision task were analysed first. The data were log transformed, and extremely fast and slow RTs were removed (RTs faster than 5.7 log-ms [300 ms] and slower than 7.8 log-ms [2500 ms]). The sleep and wake groups were initially analysed separately. A mixed-effects linear model with subjects and items as random variables, and base word condition (experimental = base words for which a new competitor was learned, control = base words for which no new competitor was learned), and time of testing (immediate, delayed, one-week later) as fixed variables was fitted for the sleep group (Figure 52, left panel). Subject-specific slopes for trial position significantly improved the fit of the model. Interaction contrasts involving base word condition and time of testing showed that the difference between the experimental and control base words was significantly different in the delayed and one-week follow up tests compared to the immediate test

(b = -0.056, t = -4.30, p < .001, and b = -0.063, t = -4.70, p < .001 respectively). No difference in this respect was found between the delayed and one-week follow up data. The difference between the two base word conditions was significant in all three time of testing conditions (immediate = 0.027, t = 2.90, p < .001, delayed: b = -0.029, t = -3.20, p < .001, one-week follow up: b = -0.036, t = -3.80, p < .001) with faster RTs to experimental compared to control base words in the immediate test, but the advantage reversing in the two latter test times. Next, the effect of time of testing was evaluated for both base word conditions separately. In the experimental base words, RTs became slower, although only marginally so, in the delayed test compared to the immediate test (b = 0.017, t = 1.80, p = .08), and speeded up significantly from the delayed to the one-week follow up test (b = -0.029, t = -3.10, p = .002) with no difference found between the immediate and the one-week follow up test. RTs to the control base words on the other hand became significantly faster from the immediate to the delayed test (b = -0.040, t = -4.30, p < .001), from the delayed to the one-week follow up test (b = -0.036, t = -3.90, p < .001), and from the immediate to the one-week follow up test (b = -0.076, t = -8.10, p < .001).

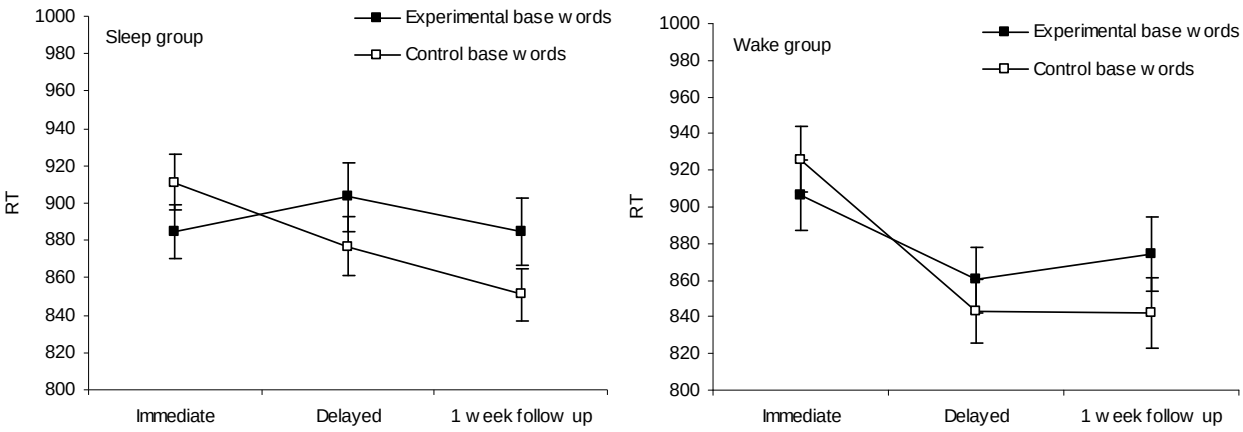


Figure 52. Lexical decision RTs to base words in sleep and wake groups. Error bars represent standard error of the means.

Accuracy rates are presented in Figure 53 (left panel), and were analysed using a mixed-effects logistic regression model with the same factors as in the RT analysis. No significant interaction contrasts were found. The simplified model showed that overall accuracy was higher in the delayed test than in the immediate test (b = 0.272, z = 2.36, p = .02[†]). No other contrasts reached significance. Although

the interaction was non-significant, visual inspection of Figure 53 suggests the advantage in the delayed test is mainly due to the control base words. Looking at the immediate vs. delayed test contrast for the control and experimental base words separately, the accuracy improvement from the immediate to the delayed test was significant for control base words only (b = 0.364, z = 2.20, p = .03†).

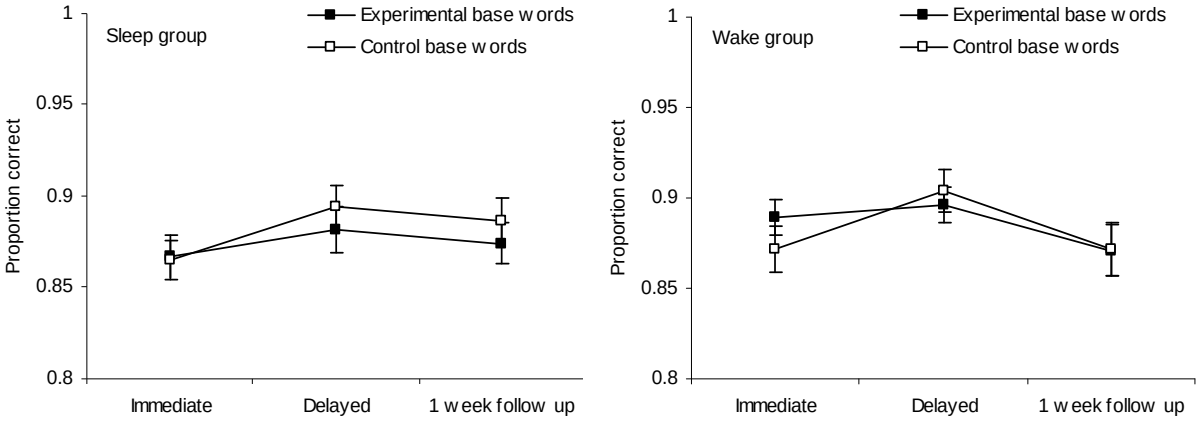


Figure 53. Lexical decision accuracy rates to base words in sleep and wake groups. Error bars represent standard error of the means.

Next the same analyses were carried out for the wake group (Figure 52, right panel). An identical model was used for this RT data set. Again, interaction contrasts showed the effect of base word condition was significantly different in the delayed and one-week follow up tests compared to the immediate test (b = -0.046, t = -3.52, p < .001, and b = -0.063, t = -4.81, p < .001 respectively). No significant difference was found in the size of the base word condition effect between the delayed and one-week follow up test. In the immediate test RTs to experimental base words were significantly faster than control base words (b = 0.024, t = 2.48, p < .001). The opposite pattern was seen in the delayed and one-week follow up tests (b = -0.021, t = -2.19, p = .036†, and b = -0.039, t = -3.90, p < .001). Looking at the effect of time of testing, RTs to experimental base words became faster from the immediate to the delayed test (b = -0.054, t = -5.84, p < .001), and the one-week follow up test (b = -0.042, t = -4.53, p < .001), but not further from the delayed to the one-week follow up test. The same pattern was seen in the control base words with significant speeding up from immediate to delayed test, but no further change from delayed to the one-week follow up (immediate vs. delayed: b = -0.099, t = -10.81, p < .001, immediate vs. one-week follow up: b = -0.106, t = -11.29, p < .001).

Accuracy rates are shown in Figure 53 (right panel). A mixed-effects logistic regression using the same factors as the RT analysis showed no significant interaction contrasts between time of testing and base word condition. The simplified model showed no difference between the two base word conditions. A significant overall increase in accuracy was found from the immediate to the delayed test ($b = 0.248$, $z = 2.09$, $p = .04^{\dagger}$), followed by a decrease in accuracy rates from the delayed to the one-week follow up test ($b = -0.360$, $z = -3.04$, $p = .002$). No significant difference was found between the immediate and the one-week follow up tests, or between the two base word conditions. Figure 53 (right panel) suggests that the change seen in accuracy rates as a function of time of testing is mainly carried by the control base words, although no interaction contrasts reached significance. Looking at the experimental and control base words separately revealed that in experimental base words there was no change from immediate to delayed test, but there was a marginally significant decrease in accuracy from the delayed test to the one-week follow up ($b = -0.323$, $z = -1.95$, $p = .05^{\dagger}$). In the control base words on the other hand there was a significant increase in accuracy from the immediate to the delayed test ($b = 0.406$, $z = 2.43$, $p = .02^{\dagger}$), followed by a significant decrease in accuracy in the one-week follow up ($b = -0.404$, $z = -2.39$, $p = .02^{\dagger}$). No difference was found between the immediate and one-week follow up tests. The difference between the two base word conditions did not reach significance in any of the test times.

Combined RT analysis of sleep and wake groups. The analyses presented above show the same pattern of data for both sleep and wake groups in terms of the difference between base word conditions in each of the three testing sessions. Both groups showed the lexical competition effect in the delayed test and in the one-week follow up. To statistically pinpoint potential differences between the two groups, data from the groups were combined in the following analysis. A mixed-effects linear model with subjects and items as random variables, and base word condition (experimental = base words for which a new competitor was learned, control = base words for which no new competitor was learned), time of testing (immediate, delayed, one-week later), and test-retest interval type (sleep vs. wake) as fixed variables was fitted. Subject-specific slopes for trial position were also added. Any differences between the sleep and wake groups would take the form of an interaction involving interval type, hence the main focus in this analysis was on such interaction

contrasts. No three-way interaction contrasts were significant, showing that the critical effect of base word condition was similar in both sleep and wake groups in all testing sessions. This interaction term was consequently dropped. The simplified model revealed a significant interaction between time of testing and interval type, showing that RTs in the wake group speeded up from the immediate to delayed and one-week tests more than in the sleep group (immediate vs. delayed: $b = -0.064$, $t = -6.90$, $p < .001$, immediate vs. one-week: $b = -0.029$ $t = -3.10$, $p < .001$). However, the contrast between the delayed session and the one-week follow up showed that the RT difference between these two times was significantly smaller in the wake group than in the sleep group ($b = 0.035$, $t = 3.70$, $p < .001$). The interaction between base word condition and interval type was non-significant, confirming that both the sleep and wake groups showed a similar lexical competition process. In sum then, there was no significant difference between the wake and sleep groups in terms of the lexical competition effect. In the immediate test both groups showed a reversed lexical competition effect, where RTs to experimental base words were facilitated relative to control base words. In the delayed test the competition effect did emerge, and was of similar magnitude in both group. It was also present in the one-week follow up, and again was of similar magnitude in both groups.

Free recall. Responses were considered accurate only if the response was made within the 3 minutes given, and if the phonetic transcription of the response completely matched the phonetic transcription of the novel word. Figure 54 shows the accuracy rates in the three test sessions (left panel). The right panel of Figure 54 shows the magnitude of change in accuracy rates between the immediate and delayed tests, and between the delayed and the one-week follow up tests. The proportion of novel words recalled accurately was analysed using a mixed effects logistic regression model, with subjects and items as random variables, and time of testing (immediate, delayed, one-week later), and test-retest interval type (sleep vs. wake) as fixed variables. In addition, it was important to assess whether the one extra exposure gained during the old/new categorisation task in the immediate test affected performance in the free recall task in the delayed test. To answer this question, old/new categorisation exposure was also added as a fixed factor in the full model (exposed vs. not exposed). If the one extra exposure resulted in increased recall rates

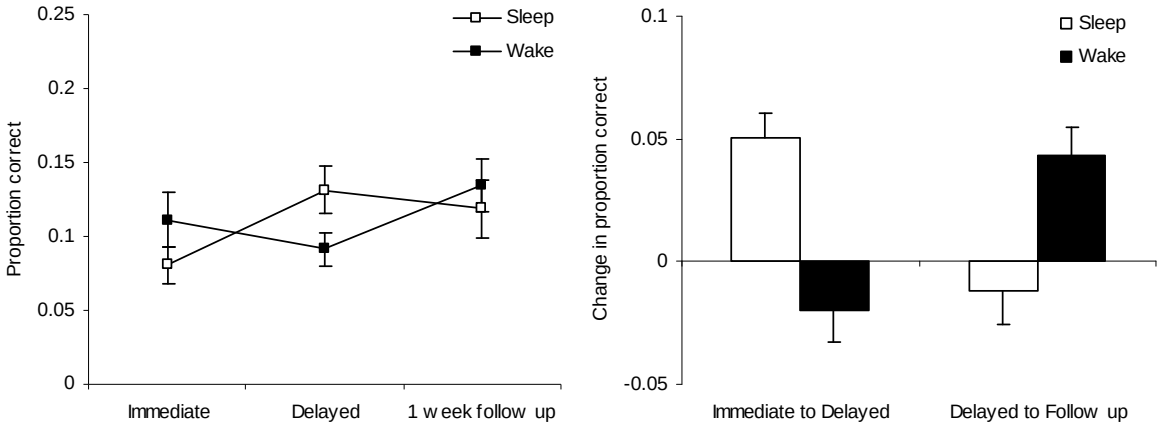


Figure 54. Accuracy rates in the free recall task, and change in accuracy rates over time. Error bars represent standard error of the means.

in the delayed test, this should be seen as an interaction between exposure and one or more of the other factors. The results showed however that exposure did not enter into interaction with any of the other factors. Hence this variable was dropped, and the subsequent analysis deals with data collapsed across the exposed and non-exposed novel words in delayed and one-week follow up sessions. In the simplified model interaction contrasts between time of testing and test-retest interval type showed that the difference in recall rates between the sleep and wake groups was significantly larger in the delayed test than either in the immediate test ($b = 0.877$, $z = 3.74$, $p < .001$) or in the one-week follow up ($b = 0.663$, $z = 2.99$, $p = .003$). Contrasts assessing the effect of interval type at each test session showed no significant difference between the sleep and wake groups in the immediate test, a marginally significant difference in the delayed test ($b = -0.468$, $z = -1.86$, $p = .06^{\dagger}$), and no difference in the one-week follow up. Next, the effect of time of testing was analysed for both the sleep and wake groups. In the sleep condition, number of words recalled increased significantly from the immediate to the delayed test ($b = 0.625$, $z = 3.78$, $p < .001$), and to the one-week follow up ($b = 0.471$, $z = 2.78$, $p = .005^{\dagger}$). The change from delayed to one-week follow up was non-significant. In the wake group on the other hand there was no significant change in recall from the immediate to the delayed test, but a significant improvement was found between the delayed and the one-week follow up tests ($b = 0.509$, $z = 3.16$, $p = .002$). The difference between the immediate and the one-week follow up did not reach significance though.

As can be seen in Figure 54, wake participants had better recall scores in the immediate test, although this difference was not statistically significant. To make sure that the changes in accuracy between the test times were not an artefact of different levels of initial performance, the data were reanalysed using subsets of participants who were matched in their recall rates in the immediate test. Participants in the wake group who had the highest accuracy rates (0.37 or higher), and participants in the sleep group who had accuracy rates of zero were removed to create subsets of 29 and 27 participants in the two groups respectively, with matched initial recall scores (0.09 in both groups). An identical model as before was fitted on the matched data, and the results revealed a nearly identical pattern of data in the matched groups. The only two changes were that the difference between the sleep and wake groups in the delayed test session now reached significance ($b = -0.613$, $z = -2.49$, $p = .01^{\dagger}$), and that in the wake group the difference between the immediate and the one-week test also now reached significance ($b = 0.358$, $z = 2.09$, $p = .04^{\dagger}$). To summarise the analysis, recall rates in the sleep group improved significantly overnight, while rates in the wake group during the day did not change. The wake group did however experience a significant improvement from the delayed to the one-week follow up test, suggesting that once sleep was allowed, a similar improvement was seen as in the sleep group. Analysis of the groups matched on initial recall confirmed this pattern was not an artefact of differences in training success.

Cued recall. The cued recall data were analysed using the same strategy as in the free recall task (Figure 55). The mixed-effects logistic regression model showed no significant interactions involving old/new categorisation exposure, hence this factor was dropped. Interaction contrasts involving time of testing and test-retest interval type showed that the difference in recall performance between the sleep and wake groups changed significantly in the delayed test ($b = -1.165$, $z = -5.95$, $p < .001$) and the one-week follow up test ($b = -0.565$, $z = -2.99$, $p = .003$) from the immediate test. The difference between the groups was non-significant in the immediate test, reached significance in the delayed test ($b = -0.717$, $z = -2.38$, $p = .02^{\dagger}$), and was non-significant in the one-week follow up. Recall performance in the sleep group improved significantly from the immediate test to the delayed test ($b = 0.534$, $z = 3.97$, $p < .001$), and from the delayed to the one-week follow up test ($b = 0.261$, $z = 2.04$, $p = .04^{\dagger}$). The difference between the immediate and the one-

week follow up tests was also significant (b = 0.797, z = 5.86, p < .001). In the wake group recall performance decayed significantly from the immediate to the delayed test (b = -0.633, z = -4.46, p < .001), but improved significantly from the delayed to the one-week follow up (b = 0.861, z = 6.23, p < .001). The improvement between the immediate test and the one-week follow up failed to reach significance.

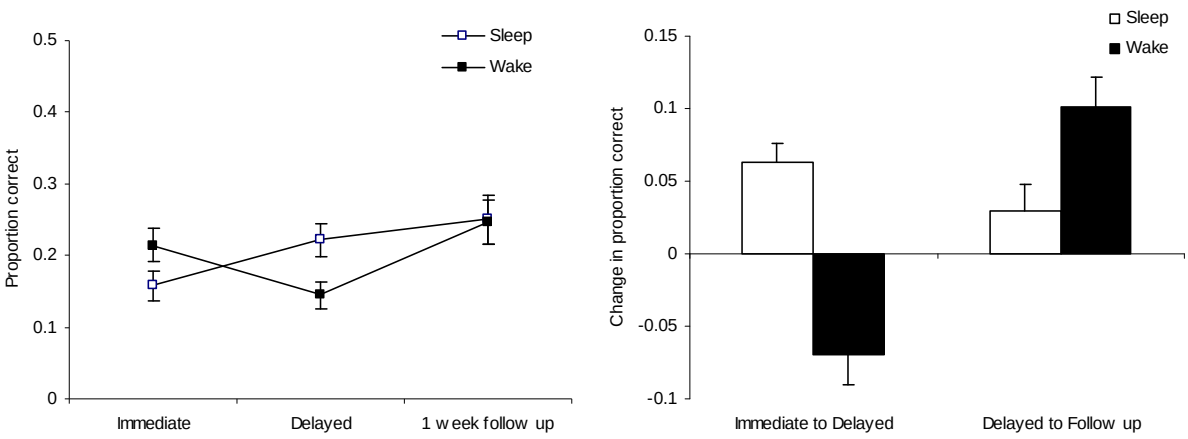


Figure 55. Accuracy rates in the cued recall task, and change in accuracy rates over time. Error bars represent standard error of the means.

As Figure 55 shows, in the immediate test there was a numerical difference between the wake and sleep groups, although this difference did not reach statistical significance. As was done in the free recall analysis, matched subsets of 26 participants in both the sleep and wake groups were selected by removing wake participants who had accuracy rates of 0.34 or higher, and sleep participants who had accuracy rates of 0.03 or lower, resulting in closely matched initial recall rates (0.17 accuracy rate in both groups in the immediate test). An identical model was fitted on these data as was used in the original analysis. The matched data showed exactly the same pattern of results as the full analysis, with one difference only. In the matched groups the difference between sleep and wake groups in the one-week follow up test now reached significance (b = -0.762, z = -2.99, p = .003), with the sleep group recalling significantly more words than the wake group. In sum, the cued recall task reflected the same main findings at the free recall task, with significant recall improvement overnight in the sleep group, but no improvement in the wake group (in fact a decline was seen here) until in the one-week follow up.

Old/new categorisation. The first analysis looked at RTs to trained novel words (Figure 56), i.e. the time it took to categorise a stimulus as a trained word (an

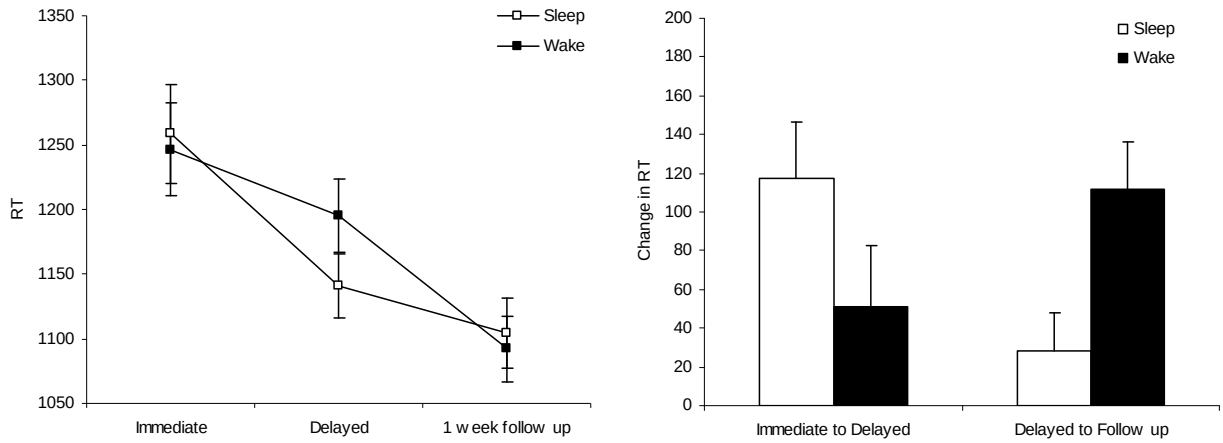


Figure 56. Response times to novel words in the old/new categorisation task, and change in RTs over time. Error bars represent standard error of the means.

"old" word). Only accurate responses were considered, and extremely fast and slow RTs were removed (RTs faster than 5.7 log-ms [300 ms] and slower than 8.0 log-ms [3000 ms]). As was done in the free and cued recall tasks, old/new categorisation exposure in the immediate test was initially included as a fixed factor to see if it modulated any of the effects associated with the other factors in the delayed and follow up tests. It did not enter into an interaction with any of the other factors, hence it was not included in the model reported here. A mixed-effects linear model with subjects and items as random variables, and test-retest interval type (sleep vs. wake) and time of testing (immediate, delayed, one-week later) as fixed variables was fitted. Subject-specific slopes for trial position significantly improved the fit of the model. Interaction contrasts between time of testing and interval type showed that the RT difference between the sleep and the wake group was significantly larger in the delayed test than either in the immediate or the one-week follow up tests ($b = -0.065$, $t = -2.94$, $p = .002$ and $b = -0.078$, $t = -4.24$, $p < .001$ respectively). The difference between the groups however was non-significant in all three test times. In the sleep group, responses became faster from the immediate to the delayed test ($b = -0.067$, $t = -4.14$, $p < .001$), and from the delayed to the one-week follow up ($b = -0.029$, $z = -2.27$, $p = .03^{\dagger}$). Similarly, the difference between the immediate and one-week follow up was significant ($b = -0.096$, $z = -5.85$, $p < .001$). In the wake group on the other hand RTs did not change significantly from the immediate to the delayed test, but there was significant improvement from the delayed to the one-

week follow up (b = -0.107, z = -8.32, p < .001). The difference between the immediate and the one-week follow up test was also significant (b = -0.109, z = -6.88, p < .001).

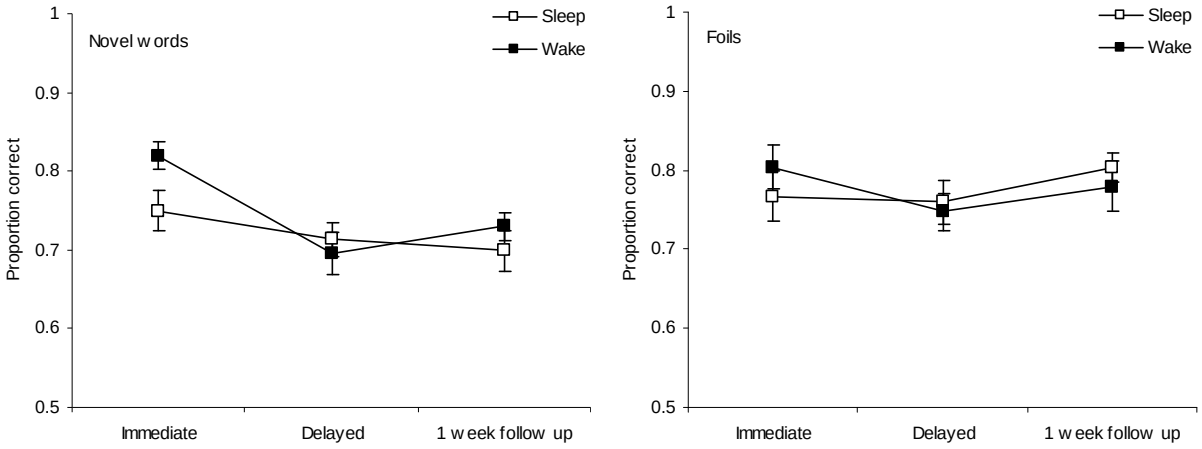


Figure 57. Accuracy rates in the old/new categorisation task. Error bars represent standard error of the means.

The accuracy rates to trained novel word trials in the old/new categorisation task are shown in the left panel of Figure 57. A mixed effects logistic regression model with subjects and items as random variables, and time of testing (immediate, delayed, one-week later), and test-retest interval type (sleep vs. wake) as fixed variables benefitted from subject-specific slopes for trial position. An interaction with time of testing and interval type showed that the accuracy difference between the sleep and wake groups was significantly different in the delayed test compared to the immediate test (b = 0.702, z = 3.34, p < .001) or to the one-week follow up test (b = 0.560, z = 3.53, p < .001). The difference between the sleep and wake groups was significant in the immediate test (b = 0.473, z = 2.11, p = .03†), failed to reach significance in the delayed test, and was marginally significant in the one-week follow up test (b = 0.331, z = 1.90, p = .057†). The effect of time of testing was assessed next for both interval groups separately. Accuracy rates in the sleep group did not change between the immediate test and the delayed test, but did decline significantly from the delayed to the one-week follow up test (b = -0.366, z = -3.27, p = .001). The difference between the immediate and the one-week follow up test was also significant (b = -0.382, z = -2.59, p = .01†). In the wake group there was a significant decline in accuracy rates between the immediate and the delayed test

(b = -0.718, z = -4.54, p < .001), followed by a non-significant improvement between the delayed test and the one-week follow up. The difference between the immediate and one-week follow up tests was significant (b = -0.524, z = -3.26, p = .001).

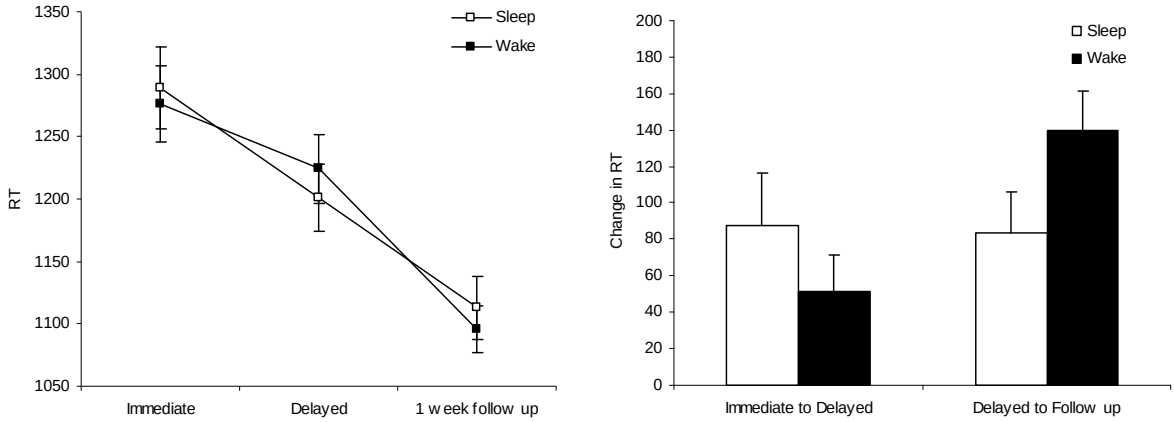


Figure 58. Response times to foils in the old/new categorisation task, and change in RTs over time. Error bars represent standard error of the means.

Categorisation times to the foils (untrained "new" words) were analysed next (Figure 58). A mixed-effects linear model with subjects and items as random variables, and interval type (sleep vs. wake) and time of testing (immediate, delayed, one-week later) as fixed variables was fitted. Subject- and item-specific slopes for time of testing significantly improved the fit of the model. No interaction contrasts reached significance. Averaged over interval type, RTs became significantly faster from the immediate to the delayed test (b = -0.062, t = -3.56, p = .001), and from the delayed to the one-week follow up test (b = -0.095, t = -6.03, p < .001), and from the immediate to the one-week follow up (b = -0.157, t = -8.12, p < .001). No difference was found between the sleep and wake groups when averaged over the test sessions. To confirm that there were no differences between the sleep and wake groups, data from both groups were analysed separately. In the sleep group responses became faster between the immediate and delayed tests (b = -0.073, t = -3.25, p = .002), between the delayed and one-week follow up tests (b = -0.076, t = -3.74, p < .001) and between the immediate and one-week follow up tests (b = -0.149, t = -5.91, p < .001). Similarly, in the wake group there was a significant improvement between the immediate and delayed test (b = -0.051, t = -2.27, p = .03†), between the delayed and one-week follow up test (b = -0.114, t = -5.57, p < .001) and between the immediate and one-week follow up test (b = -0.165, t = -6.55, p < .001). No

significant difference between the sleep and wake groups was found at any of the three testing times.

Accuracy rates in categorising foils (untrained "new" words) are presented in the right panel of Figure 57. A mixed effects logistic regression model with subjects and items as random variables, and time of testing (immediate, delayed, one-week later), and interval type (sleep vs. wake) as fixed variables benefitted from subject- and item-specific slopes for trial position. Interaction contrasts showed that there was a significant change in the difference between the wake and sleep groups from the immediate test to the delayed ($b = -0.453$, $z = -2.01$, $p = .04^{\dagger}$) and one-week follow up tests ($b = -0.647$, $z = -2.81$, $p = .005^{\dagger}$), reflecting the wake advantage in the immediate test changing into a sleep advantage in the two later tests. There however was no significant difference between the sleep and wake groups in any of the three test sessions. Looking at the effect of time of testing in the two interval groups separately, the model showed that in the sleep group there was a significant improvement in accuracy from the delayed to the one-week follow up test ($b = 0.543$, $z = 3.83$, $p < .001$). No other contrasts reached significance in this group. In the wake group there was a significant decline in accuracy from the immediate to the delayed test ($b = -0.686$, $z = -3.96$, $p < .001$), but a significant improvement from the delayed to the one-week follow up ($b = 0.351$, $z = 2.61$, $p = .009^{\dagger}$). The difference between the immediate and the one-week follow up tests was marginally significant ($b = -0.336$, $z = -1.93$, $p = .05^{\dagger}$).

In the old/new categorisation task the most critical condition was the one where categorisation responses were made to the novel words ("old") as this gives a measure of recognition time directly to the novel words. Here sleep group RTs improved overnight, while no significant change was seen in the wake group. The wake group did improve by the one-week follow up, and an improvement was seen in the sleep group as well. In the one-week follow up RTs in the two groups were nearly identical.

6.2.2.2 Self-reported measures of alertness

Mean scores from the alertness and sleepiness questionnaire in each test session are presented in Table 5. An ordinal logistic regression model was used to analyse the Stanford Sleepiness Scale data, with test-retest interval type (sleep vs. wake) and time of testing (immediate, delayed, one-week follow up) as predictors.

These two factors did not enter into an interaction, hence the term was dropped. Averaged across the time of testing conditions, there was no significant difference between the sleep and wake groups. There was a significant change in the scores from the immediate to the delayed test ($b = -2.783$, $z = -7.68$, $p < .001$), but no further change from the delayed to the one-week follow up. The difference between the immediate and one-week follow up was significant ($b = -3.197$, $z = -7.68$, $p < .001$). Although there was no significant interaction between the two factors, Table 5 suggests there was a numerical difference between the sleep and wake groups at least in the immediate test. This difference however did not reach significance in any of the three testing sessions. When examined individually, both the sleep and wake groups showed the same pattern of change over time as was seen in the overall analysis above (for sleep group immediate vs. delayed: $b = -3.152$, $z = -5.86$, $p < .001$, delayed vs. follow up: $b = -0.667$, $z = -1.35$, $p = .18$, immediate vs. follow up: $b = -3.819$, $z = -6.79$, $p < .001$, for the wake group immediate vs. delayed: $b = -2.496$, $z = -4.83$, $p < .001$, delayed vs. follow up: $b = -0.181$, $z = -0.37$, $p = .71$, immediate vs. follow up: $b = -2.676$, $z = -5.20$, $p < .001$).

Table 5. Self-reported measures of sleepiness and alertness in the sleep and wake groups.

		Immediate test	Delayed test	One-week follow up
Stanford scale				
	Sleep	3.90 (0.23)	2.13 (0.14)	1.87 (0.13)
	Wake	3.45 (0.18)	2.26 (0.19)	2.07 (0.14)
Ability to concentrate				
	Sleep	5.04 (0.39)	8.22 (0.27)	9.02 (0.29)
	Wake	6.79 (0.36)	8.54 (0.34)	8.81 (0.28)
Feeling refreshed				
	Sleep	4.16 (0.32)	8.38 (0.36)	8.61 (0.34)
	Wake	5.26 (0.38)	7.74 (0.43)	8.19 (0.36)

Note. Standard error in parentheses. In the Stanford scale 1 = awake, 7 = sleepy. In the visual analogue scales 0 = poor/not at all refreshed, 11.5 = excellent/very refreshed.

The data from the visual analogue scale asking participants to rate their ability to concentrate were analysed next (Table 5). Here participants were asked to mark the printed line at a location that best corresponded to their current ability to concentrate. The location of the mark was measured (in centimetres) from the beginning of the line, giving a numerical value between 0 and 11.5, where 0 corresponds to “poor” and 11.5 to “excellent”. A mixed-effects linear model with subjects as random factors, and test-retest interval type (sleep vs. wake) and time of

testing (immediate, delayed, one-week follow up) as fixed factors was used to analyse the data. Interaction contrasts showed that the difference between the sleep and wake groups was significantly smaller in the delayed ($b = -1.426$, $t = -2.70$, $p = .008^{\dagger}$) and one-week follow up tests ($b = -1.951$, $t = -3.65$, $p < .001$) compared to the immediate test. No difference in the magnitude of this effect was seen between the delayed and one-week follow up tests. The difference between the sleep and wake groups was significant in the immediate test ($b = 1.748$, $t = 3.79$, $p < .001$), with participants in the wake condition reporting better level of concentration. This difference however was non-significant in all subsequent test sessions. In the sleep group there was a significant increase in ability to concentrate from the immediate to the delayed test ($b = 3.174$, $t = 8.49$, $p < .001$) and from the delayed to the one-week follow up test ($b = 0.812$, $t = 2.15$, $p = .03^{\dagger}$), as well as from the immediate to the one-week follow up test ($b = 3.986$, $t = 10.55$, $p < .001$). A similar pattern was seen in the wake group, with concentration scores increasing from the immediate to the delayed test ($b = 1.748$, $t = 4.68$, $p < .001$), although not from the delayed to the one-week follow up. The difference between the immediate and the one-week follow up tests was significant ($b = 2.035$, $t = 5.39$, $p < .001$).

An identical linear model was used to analyse data from the scale asking participants to rate how refreshed they felt. Here 0 corresponded to “not at all refreshed”, and 11.5 corresponded to “very refreshed”. Interaction contrasts showed that the difference between the sleep and wake groups was significantly smaller in the delayed ($b = -1.742$, $t = -2.79$, $p = .006^{\dagger}$) and in the one-week follow up test ($b = -1.528$, $t = -2.43$, $p = .02^{\dagger}$) compared to the immediate test. The difference between delayed and follow up tests was non-significant. Further inspection of contrasts showed that the difference between the sleep and wake groups was significant in the immediate test ($b = 1.100$, $t = 2.13$, $p = .03^{\dagger}$), but non-significant in the following two test sessions. Looking at the ratings as a function of time of testing, in the sleep group ratings increased from the immediate to the delayed test ($b = 4.223$, $t = 9.58$, $p < .001$), but did not change from the delayed to the follow up test. The difference between the immediate and the follow up test was significant ($b = 4.453$, $t = 10.00$, $p < .001$). The same pattern was seen in the wake group, with a significant increase in ratings from the immediate test to the delayed test ($b = 2.481$, $t = 5.63$, $p < .001$), but no further increase from the delayed to the one-week follow

up. The difference between the immediate and the follow up tests was significant (b = 2.926, t = 6.57, p < .001).

The above data can be summarised by drawing attention to two findings. The first finding, supported by all three scales, was that participants were more tired and less attentive in the immediate test than in either of the two later tests. This was true of both the sleep and the wake group. The effect was probably caused by the presence of the training session which was likely to cause fatigue. However, as the same effect holds for both groups, it cannot account the sleep vs. wake differences reported above. The second finding, supported only by the two visual analogue scales, was that the sleep group was less attentive in the immediate test than the wake group. This may explain why there was a trend in accuracy rates in the free and cued recall tasks in favour of the wake group. The analyses in those tasks using groups matched on immediate recall rates however showed that this initial difference did not affect the subsequent performance differences between the groups in the two delayed test sessions. Hence it seems that the changes in self-reported fatigue and attentiveness cannot account for the differences between the sleep and wake groups. It is also difficult to see how these measures could explain the delayed emergence of the lexical competition effect, although the possibility that fatigue blocks the competition effect cannot be conclusively ruled out based on these data. However, in the light of several earlier reports of a delayed competition effect (Gaskell & Dumay, 2003; Dumay et al., 2004; Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008; Davis et al., 2009) or a strengthening competition effect over time (Fernandes et al., 2009), the fatigue explanation can be regarded as unlikely. There was also no consistent evidence for circadian effects in the questionnaire data: these should manifest in differences between the sleep and wake groups. The only difference was found in the immediate test session. Circadian effects might also be seen in the test performance data with performance improving in the evening compared to the morning (Hasher et al., 2005). Such a pattern however was not seen in any of the tests.

6.2.2.3 Polysomnographic measures

Sleep stages. Pearson correlation coefficients were calculated for the correlation between the word learning measures in the four tasks and total sleep time, time spent in stage 2 sleep, in SWS, and in REM sleep. The word learning measures

Table 6. Correlations between word learning measures and time spent in different sleep stages.

			TST	Stage 2	REM	SWS
Lexical competition						
	Change overnight	*r*	-0.165	-0.056	0.069	-0.191
		p	*0.38*	*0.77*	*0.71*	*0.30*
	Immediate test	*r*	0.338	0.128	0.068	0.184
		p	*0.06*	*0.49*	*0.71*	*0.32*
	Delayed test	*r*	0.153	0.069	0.187	-0.072
		p	*0.41*	*0.71*	*0.31*	*0.70*
	Follow up test	*r*	-0.181	-0.084	-0.143	0.057
		p	*0.34*	*0.67*	*0.45*	*0.77*
Free recall						
	Change overnight	*r*	0.108	0.256	0.038	-0.271
		p	*0.56*	*0.16*	*0.84*	*0.14*
	Immediate test	*r*	-0.182	-0.131	-0.116	0.11
		p	*0.33*	*0.48*	*0.53*	*0.56*
	Delayed test	*r*	-0.072	0.063	-0.071	-0.079
		p	*0.70*	*0.74*	*0.70*	*0.67*
	Follow up test	*r*	-0.326	-0.146	-0.289	-0.03
		p	*0.08*	*0.44*	*0.12*	*0.87*
Cued recall						
	Change overnight	*r*	0.072	-0.008	0.064	0.134
		p	*0.70*	*0.97*	*0.73*	*0.47*
	Immediate test	*r*	-0.239	0.037	**-0.364**	-0.02
		p	*0.20*	*0.84*	***0.044***†	*0.91*
	Delayed test	*r*	-0.164	0.03	-0.283	0.062
		p	*0.38*	*0.87*	*0.12*	*0.74*
	Follow up test	*r*	-0.292	-0.082	**-0.372**	-0.042
		p	*0.13*	*0.67*	***0.047***†	*0.83*
Old/new categorisation	Change overnight	*r*	-0.034	**-0.385**	-0.117	**0.495**
		p	*0.86*	***0.033***†	*0.54*	***0.005***†
	Immediate test	*r*	-0.165	-0.345	0.087	0.250
		p	*0.37*	*0.06*	*0.65*	*0.18*
	Delayed test	*r*	-0.215	-0.102	-0.003	-0.169
		p	*0.25*	*0.59*	*0.99*	*0.36*
	Follow up test	*r*	-0.015	-0.087	0.184	0.012
		p	*0.94*	*0.65*	*0.34*	*0.95*

Note: Significant correlations in bold. TST = total sleep time, REM = time spent (in minutes) in rapid eye movement sleep, SWS = time spent (in minutes) in slow wave sleep. † = p-values that do not survive a Bonferroni correction for multiple comparisons.

included the magnitude of the lexical competition effect (difference in RTs to experimental and control base words), free and cued recall performance, and old/new categorisation RTs to trained novel words. The correlation coefficients and their corresponding p-values are presented in Table 6. The p-values in the table are

uncorrected for multiple comparisons, however Bonferroni adjusted p-values were calculated (with alpha level at .05) and those tests that failed to meet the adjusted alpha level are marked with the symbol †. The correction was calculated by considering contrasts within each test task as multiple comparisons. The rows titled "Change overnight" refer to the difference in performance between the immediate and the delayed test. The table shows that the strongest correlation was between time spent in SWS and the difference in old/new categorisation RTs from the immediate to the delayed test, with increasing SWS associated with larger RT improvement overnight. The opposite pattern was seen with stage 2 sleep duration, with larger improvement associated with decreasing stage 2 duration. This suggests that in the participants whose categorisation RTs improved the most the increase in SWS may have been at the expense of stage 2 sleep. There were also statistically less reliable correlations between time spent in REM and cued recall performance in the immediate and one-week follow up tests, with increasing REM durations associated with poorer cued recall performance.

Another common way of analysing sleep stage correlations is to look at percentage of REM, SWS and stage 2 sleep rather than the actual time spent in these stages. The advantage of this analysis is that it focuses on effects of sleep stages independent of total sleep time. This analysis showed similar results as the analysis in Table 3. The correlation between SWS and old/new categorisation change overnight was still highly significant ($r = 0.483$, $p = .006^{\dagger}$), as was the correlation with stage 2 sleep ($r = -0.462$, $p = .009^{\dagger}$). The correlations between cued recall and REM were no longer significant. This lack of significance in the proportional analysis and the high p-values (which did not reach corrected significance) in the non-proportional analysis suggest that the cued recall correlations may be unreliable and should be viewed with caution. The old/new categorisation correlations also failed to reach corrected significance, but did reach uncorrected significance in both analyses, suggesting that they are more reliable.

Sleep spindles. The sleep spindle analysis included stage 2 sleep only, as this is the sleep stage where great majority of sleep spindles occur. The first step in data processing was to remove any epochs containing wake or brief arousals, as well as artefacts caused by movements and poor recording quality from one or more electrodes. Any electrodes that consistently provided a noisy signal were removed from the analysis. The raw EEG data were then band-pass filtered between 11 and 15

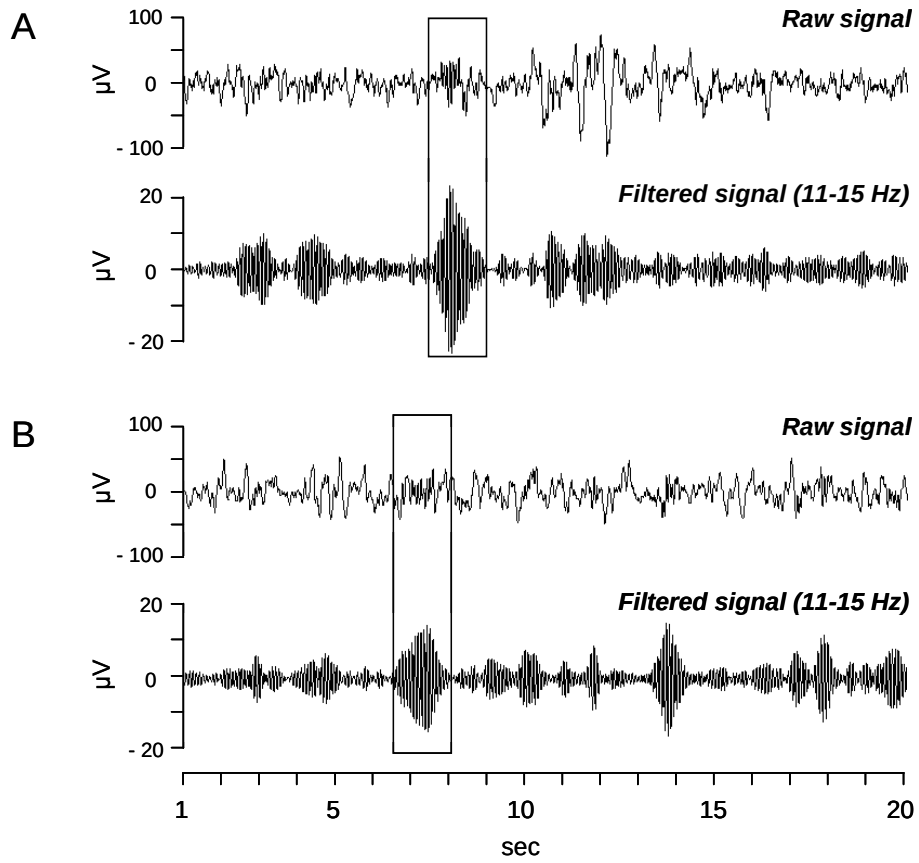
Hz using a linear finite impulse response (FIR) filter provided in the EEGLAB toolbox for Matlab (Delorme & Makeig, 2004). An automated EEG spindle detection algorithm, developed by Ferrarelli et al. (2007) and implemented in Matlab, was then used to derive the spindle measures (number and amplitude of spindles). This algorithm uses the amplitude of the filtered signal to generate a time series for each channel. Any amplitude fluctuation that exceeds a pre-determined upper threshold is counted as a spindle, with the peak amplitude for each spindle defined as the local maximum above the threshold. The beginning and end of a spindle were defined as amplitudes preceding and following the peak, up to a point where the amplitude crossed a lower threshold. The upper and lower thresholds were calculated relative to the mean signal amplitude in the channel, as the mean amplitude varies across channels. The lower and upper thresholds were set at two and eight times the average amplitude. These values were selected by Ferrarelli et al. (2007) to give a good match with visually detected spindles. This algorithm has been reported by both Ferrarelli et al. (2007) and by Nishida and Walker (2007) to give reliable spindle counts when compared with visual scoring, although it appears to be fairly conservative, as discussed in Figure 59.

Table 7. Sleep spindle measures at each electrode.

	Total N	Density	Ampl.
C3	269 (25)	0.50 (0.04)	22.47 (0.87)
C4	246 (22)	0.46 (0.04)	19.75 (0.87)
F3	316 (23)	0.59 (0.04)	19.68 (0.92)
F4	297 (20)	0.55 (0.04)	16.28 (0.82)

Note: Standard error in parentheses. Total N = total number of spindles, Density = spindle density (number of spindles per 30 seconds), Ampl. = mean maximal spindle amplitude in μV.

Table 7 shows the most often used measures of spindle activity at each of the four electrodes. These measures consist of total number of detected spindles, spindle density (average number of spindles per 30 s of stage 2 sleep), and mean maximal spindle amplitude (average of the highest amplitude points for each detected spindle). A mixed-effects linear model with subjects as random factor, and electrode site as fixed factor showed that electrodes on the left recorded a greater number of spindles than electrodes on the right (C3 vs. C4: b = -23.680, t = -2.87, p = .006†, F3 vs. F4: b = -17.535, t = 2.12, p = .03†). The same was true of spindle density, with a higher density on the left than on the right (C3 vs. C4: b = -0.041, t = -2.66, p = .01†,



The figures above show two 20-second epochs of stage 2 sleep at different times from the same participant and channel. Panel A identifies a spindle detected by the algorithm used in the current experiment. Panel B highlights an event that would be likely classified as a spindle by visual inspection, but which the automatic detector did not pick up. Visual detection is typically based on criteria relating to frequency, duration, and amplitude. For example, Fogel and Smith (2006) categorised as spindles events occurring during stage 2 sleep that were in the 12-16 Hz range, exceeded 0.5 s in duration, and had a maximal amplitude of at least 10 µV. Many automatic detection algorithms have also used similar fixed criteria (e.g., Schabus et al., 2004). The current algorithm uses a variable amplitude threshold based on the average amplitude of the channel in question. This is because amplitude can vary significantly between channels even within a participant, making a fixed amplitude threshold vulnerable to false positives. The downside is that the algorithm may miss spindles particularly in noisy channels as in these cases the threshold may end up being quite high. Comparison of the spindle densities calculated with this algorithm and those reported in earlier studies indeed show that the algorithm is conservative. However, as this will be the case for each participant, the comparisons between participants reported here are unaffected.

Figure 59. Examples of spindles detected and missed by the automatic detection script.

F3 vs. F4: b = -0.032, t = 2.06, p = .04†), and of mean spindle amplitude with spindles detected by the left electrodes having a larger amplitude (C3 vs. C4: -2.785, t = -11.99, p < .001, F3 vs. F4: b = -2.822, t = -12.15, p < .001). Furthermore, the frontal electrodes were associated with a higher total number of spindles (C3 vs. F3: b = 59.065, t = 7.15, p < .001, C4 vs. F4: b = 65.211, t = 7.90, p < .001), higher spindle density (C3 vs. F3: b = 0.112, t = 7.16, p < .001, C4 vs. F4: b = 0.121,

$t = 7.77$, $p < .001$), and higher maximal amplitude (C3 vs. F3: $b = 2.515$, $t = 10.83$, $p < .001$, C4 vs. F4: $b = 2.552$, $t = 10.99$, $p < .001$).

Table 8 shows correlations between the novel word learning measures and two of the spindle activity measures: spindle density and maximal spindle amplitude, averaged over the four electrodes (see Appendix 11 for correlations for each electrode site individually). Total number of spindles was left out of the analysis to reduce the number of multiple correlations as it was highly correlated with spindle density ($r = 0.915$, $p < .001$), confirming that both measure effectively the same variable[6]. As above, those tests that fail to meet the Bonferroni adjusted alpha level are marked with the symbol †. The row titled "Change overnight" again refers to the difference between performance in the delayed and the immediate test sessions. The only aspect of these data that was associated with spindle activity was lexical competition. In the immediate test, participants who showed the least lexical competition (or the most facilitation) went on to experience most spindle activity during the following night (Figure 60). This correlation was seen in both measures of spindle activity (spindle density: $r = -0.536$, $p = .002$, maximal spindle amplitude: $r = -0.389$, $p = .027$†), the density measure was reliable at all four electrode sites (see Appendix 11). Furthermore, spindle activity correlated with the magnitude of change in the lexical competition effect overnight, with increasing competition effect being associated with higher spindle activity (Figure 61). This was the case spindle density ($r = 0.599$, $p < .001$) at all four electrode sites (Appendix 11), although the correlation with maximal amplitude did not reach significance in the analysis.

Slow vs. fast spindles. The spindle correlations were analysed also for slow and fast spindles separately. These were counted by band-pass filtering the artefact-rejected data between 11 and 13 Hz for slow spindles, and between 13 and 15 Hz for fast spindles (following Schabus et al., 2007). The automatic spindle counting algorithm was then applied on both data sets. The data were broadly in agreement with the overall analysis. No significant correlations were found with free recall, cued recall, or old/new categorisation. The magnitude of the lexical competition effect in the immediate test correlated with both measures of fast spindle activity (density: $r = -0.637$, $p < .001$, amplitude: $r = -0.434$, $p = .015$†), but only with slow

[6] Same pattern of results is seen if total number of spindles is considered instead of spindle density.

Table 8. Correlations between word learning measures and sleep spindle activity (11-15 Hz).

			Density	Ampl.
Lexical competition				
	Change overnight	*r*	**0.599**	0.306
		p	***< 0.001***	*0.094*
	Immediate test	*r*	**-0.536**	**-0.398**
		p	***0.002***	***0.027***†
	Delayed test	*r*	0.275	0.008
		p	*0.14*	*0.96*
	Follow up test	*r*	0.121	0.139
		p	*0.53*	*0.46*
Free recall				
	Change overnight	*r*	-0.188	-0.119
		p	*0.31*	*0.53*
	Immediate test	*r*	-0.243	0.031
		p	*0.19*	*0.87*
	Delayed test	*r*	-0.306	-0.042
		p	*0.09*	*0.82*
	Follow up test	*r*	-0.307	-0.162
		p	*0.10*	*0.39*
Cued recall				
	Change overnight	*r*	-0.05	0.09
		p	*0.79*	*0.63*
	Immediate test	*r*	-0.303	-0.056
		p	*0.10*	*0.76*
	Delayed test	*r*	-0.29	0.003
		p	*0.11*	*0.99*
	Follow up test	*r*	-0.243	-0.191
		p	*0.20*	*0.32*
Old/new categorisation	Change overnight	*r*	-0.126	0.143
		p	*0.50*	*0.44*
	Immediate test	*r*	-0.008	0.031
		p	*0.97*	*0.87*
	Delayed test	*r*	0.128	-0.113
		p	*0.49*	*0.55*
	Follow up test	*r*	0.215	-0.055
		p	*0.26*	*0.77*

Note: Significant correlations in bold. Density = spindle density (number of spindles per 30 seconds), Ampl. = average maximal spindle amplitude. † = p-values that do not survive a Bonferroni correction for multiple comparisons.

spindle density (r = -0.410, p = .022†). The overnight change in the lexical competition effect correlated with both measures of fast spindle activity (density: r = 0.664, p < .001, amplitude: r = 0.369, p = .041†), and with slow spindle density (r = 0.479, p = .006).

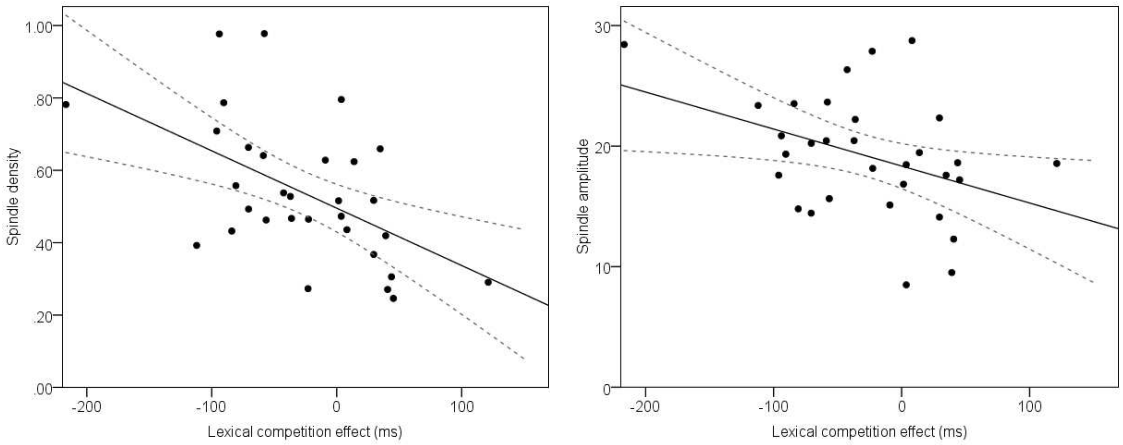


Figure 60. Scatterplots showing correlations between lexical competition effect immediately after training and spindle activity during the subsequent night. Dashed lines represent 95% confidence intervals. The x-axis shows the magnitude of the lexical competition effect (RTs to experimental base words – RTs to control base words), where positive values indicate lexical competition and negative values indicate a facilitatory effect.

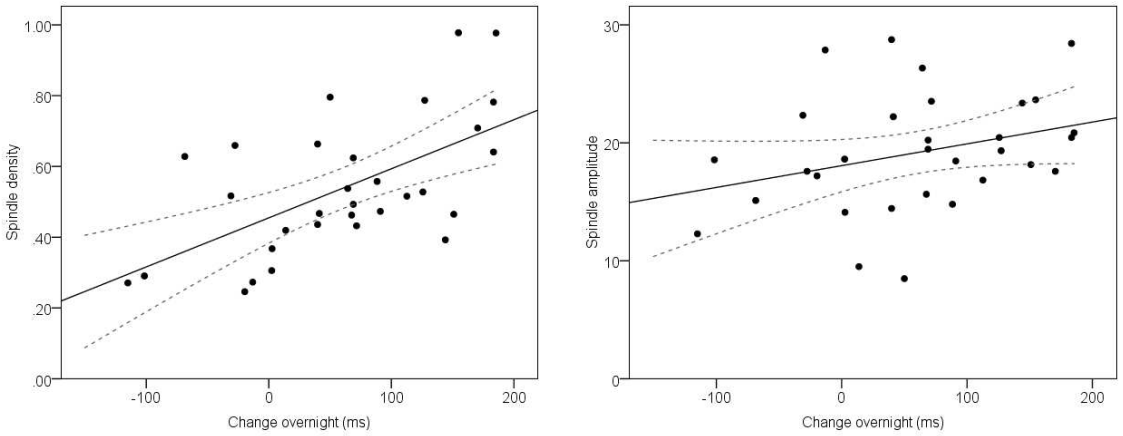


Figure 61. Scatterplots showing correlations between change in lexical competition effect overnight and spindle activity. Dashed lines represent 95% confidence intervals. The x-axis shows the change in the magnitude of the lexical competition effect from the immediate test to the first delayed test after a night of sleep.

The correlations presented in Table 8 suggest that spindle activity is associated with the degree of lexical competition immediately after training, whereby those participants who show the least evidence for competition experience more spindle activity during the following night. The table also highlights a correlation with the magnitude of change in the competition effect, whereby participants who experience a larger increase in competition overnight also show higher spindle activity. It is however possible that both spindle correlations have a

common source. This might be the case if the two behavioural measures, immediate competition and overnight change in competition, were also associated. This was tested by calculating the correlation between the two behavioural measures. The correlation was highly significant ($r = -0.758$, $p < .001$), showing that weak initial competition (or in other words strong facilitation) was associated with a large increase overnight in the competition effect. In light of this relationship, it was important to re-evaluate the correlation between competition change and spindle activity, while controlling for the association with the immediate competition effect. A partial correlation was calculated between overnight competition change and spindle density, while holding immediate competition constant. Here the correlation was marginally significant ($r = 0.350$, $p = .058$)[7], suggesting that while spindle activity is associated with immediate performance, it also makes a marginally significant contribution to the competition change overnight independent of immediate performance.

6.3 Chapter Summary and General Discussion

The experiment reported in this chapter had two primary aims. The first was to see if offline consolidation of newly learned words occurs preferentially during sleep compared to wake. This was examined both in terms of explicit recall and recognition of the novel words themselves, and in terms of integrating the novel words in the mental lexicon. The second aim was to identify those aspects of sleep architecture (if any) that are associated with overnight consolidation of novel words.

The two tasks measuring explicit recall of novel words consisted of free recall and cued recall. In free recall participants who spent the first test-retest interval asleep (sleep group) recalled 5.0% words more in the morning than they did immediately after training. This was a statistically significant improvement. Participants who remained awake for an equivalent time (wake group) on the other hand showed a 1.2% non-significant decline in recall. Interestingly, when tested again about one week later, the recall rates in the wake group improved significantly by 4.3% from the delayed test, while recall in the sleep group did not improve any further (a non-significant decline of 2% was seen). As seen in Figure 54, in this last

[7] The same correlation reaches statistical significance if total number of spindles rather than spindle density is considered ($r = 0.381$, $p = .038$).

test session both groups showed equally good recall. These data suggest that both groups eventually reached an equivalent level of recall, and both groups improved from the immediate test to the one-week follow up test, but the timing of sleep determined the point when the improvement was seen. The observation that the improvement followed sleep in the sleep group but was not seen after a similar delay of wakefulness in the wake group suggests that the improvement was a result of sleep-dependent consolidation.

A roughly similar pattern was seen in cued recall. Here sleep participants improved overnight by a significant 6.3%, while the wake group's recall accuracy declined significantly over the course of the day, by 7.0%. However, the wake group's accuracy rose by 10.2% between the delayed and the one-week follow up tests (recall also that these two tests took place at the same circadian time), while the sleep group also experienced a smaller but statistically significant further increase of 2.9%. The cumulative effect of these changes was that by the end of the experiment the sleep group's recall rates had improved significantly from the immediate test, while the wake group's had not (because of the initial decline), suggesting that in this task in particular the timing of sleep played a crucial role, with immediate sleep allowing greater gains in performance over time. This is an important point as it clarifies the effect seen in Experiment 5, where participants in a visual cued recall task recalled more words that were learned on the previous day compared to words learned on the day of testing. That design did not allow a distinction between explanations of time-dependent and sleep-dependent offline consolidation, as the interval between learning and testing contained both sleep and wake. The current data however suggest that sleep provides the optimal environment for consolidation in this task.

Finally, the old/new categorisation task was used as a task measuring speed of novel word recognition. The sleep group showed similar advantage here as in the recall tasks. RTs to novel words in the sleep group became significantly faster overnight, with a 117 ms improvement, while the wake group showed an improvement only half as large as the sleep group, with a non-significant 51 ms. The sleep group went on to show a smaller but significant further improvement of 28 ms from the delayed to the one-week follow up test. The wake group on the other hand showed a large and significant improvement of 112 ms between the delayed and the follow up test. The cumulative result was that both groups had improved

significantly from the immediate to the one-week follow up test, and did not differ in RTs at the last test session, but again timing of sleep determined in which test session the improvement was seen. The accuracy rates to novel words supported a sleep advantage by showing no change in accuracy overnight, but a significant decline over the course of a day. In the sleep group accuracy rates declined by the one-week follow up, while the wake group improved. However, compared to immediate performance, both groups had declined by the follow up a week later. This suggests that some of the RT gains may have been due to a speed-accuracy trade off, however this would have been the case in both the sleep and the wake group, thus the differences between the groups cannot be attributed to this.

The current data are in agreement with the free recall data reported by Dumay and Gaskell (2007), who found that recall increased in the sleep group overnight, while there was a trend towards a decline in performance in the wake group over the course of a day. After a further 12 hours the sleep group's recall rates increased still slightly but significantly, while the wake group now experienced a large and significant improvement, probably because this group had now also had a chance to sleep. The data from Experiment 8 are very similar, although the sleep group here did not further improve between the delayed and one-week follow up in free recall. The current cued recall data pattern however was very similar to Dumay and Gaskell's free recall data. The current experiment also improved upon Dumay and Gaskell's 2AFC task by using a modified version that allowed recording of recognition times. Dumay and Gaskell did not find a sleep advantage in this task, but Davis et al. (2009) did show higher accuracy in response to words learned a day before testing, compared to words learned on the day of testing. The data from the current experiment suggest this effect may have been sleep-related, with improving RTs overnight, and a sleep advantage in accuracy rates.

The lexical competition results in the current experiment did not fully replicate the pattern seen by Dumay and Gaskell (2007). While Dumay and Gaskell found lexical competition effects only after a night of sleep, here the effect emerged in the delayed test in both sleep and wake groups. While numerically the competition effect was smaller in the wake group (27 ms in sleep group, 17 ms in wake group), the analysis combining the two groups did not show a significant difference in the magnitude of the effect between the groups. In contrast, nearly identical lexical competition effects were found for both groups in the one-week follow up (33 ms for

both sleep and wake groups), as well as in the immediate test (-26 ms in the sleep group, -20 ms in the wake group). The current data showed an initial facilitatory effect whereby experimental base words were responded to faster than control words. Dumay and Gaskell reported no difference between the conditions in their immediate test.

There are several factors that may contribute to the discrepancy in the time course of the emergence of lexical competition effects between the current experiment and that of Dumay and Gaskell. These studies differed both in the stimuli they used, and the task used to measure lexical competition. Dumay and Gaskell used pause detection, where a short (200 ms) period of silence was inserted towards the end of a base word (e.g., *cathedr_al*). The time it takes to make a pause detection decision is taken as a measure of lexical activity at that point in time, with high degree of lexical competition resulting in slower pause detection times (Mattys & Clark, 2002). Hence base words for which a new competitor had been acquired were associated with slower pause detection times compared to base words for which no new competitor had been trained. The advantage of this task is that it does not require participants to make explicit decisions about the identity of the base word. It may be the case that in the lexical decision task some participants choose a more strategic approach and delay their responses in order to make sure they do not confuse base words with novel words. Such an effect would probably not depend on sleep specifically, and would mask a sleep benefit. Although this latter explanation is made less likely by the fact that in the current experiment novel words were never presented in the lexical decision task, it may still be that pause detection as a completely implicit gauge of lexical activity provides a purer and more sensitive measure of lexical competition. It should be noted though that the shift from facilitation in the immediate test to competition in the delayed test would be difficult to explain in the context of a strategic lexical competition effect which should be evident irrespective of time of testing. The initial facilitation seen here may be due to phonological priming effects carrying over from repeated exposure to similar sounding novel words in training. Pause detection is less likely to be affected by such priming (as no overt recognition response is required).

Another important difference concerned the nature of the stimuli. Dumay and Gaskell generated their novel words by adding a consonant cluster at the end of a base word (e.g., *shadowks*), while the current experiment changed the final

phonemes of a base word (e.g., *cathedruke*). This difference is a likely explanation for why free recall rates in the current experiment were much lower than in Dumay and Gaskell (accuracy rates in Figure 54 vary between 8% and 13%, while Dumay and Gaskell's Figure 1 suggests variation between 15% and 30%). Recalling a familiar word with a novel added ending may be easier than recalling a novel word that is a variation of a familiar word, although this should not affect the time course of lexical integration.

Finally, the interpretation of the emergence of the lexical competition effect in the wake group is complicated by the fact that there was no objective control over what the wake participants did during the course of the day. Although they were asked not to consume stimulating substances, or to sleep, compliance was not verified by objective measures. Hence it is possible that participants varied in terms of the external stimulation they underwent during the day. For example, if a large proportion of participants napped during the day, this might have been enough to bring out statistically reliable lexical competition effects. The likelihood of this latter possibility can be evaluated on the basis of the information collected about participants' sleep habits. In fact, when the data from the wake group were reanalysed including only those participants who did not habitually tend to nap during the day (n = 21), the lexical competition effect in the 10-hour delayed test no longer reached significance. It did still reach significance in the one-week follow up. The facilitatory effect in the immediate test also failed to reach significance in this sub-set of participants, suggesting that perhaps the change might have been partially due to reduced statistical power, as there is no reason why habitual nappers would show a different pattern of performance in the immediate test from non-nappers. The reanalysis however reinforces the view that future studies should control the wake group's environment more carefully, either requiring them to remain in the laboratory, or by means of actigraphy. This is a method where participants wear a non-invasive actimetry sensor which monitors body movements and allows identification of long periods of lack of movement, indicating sleep.

It is important here also to address the issue of potential circadian effects. The recall and old/new categorisation improvements seen in the sleep group overnight, and the corresponding lack of improvements seen in the wake group might be due to the fact that the delayed test was carried out at different times of day, with the sleep group doing the test in the morning and the wake group doing it in the

evening. Hence the wake group may have been more tired at test, and perform poorly. This alternative explanation can be assessed in two ways. Firstly, if the wake group were more tired in the delayed test, this should have been reflected in the participants' subjective evaluations of alertness. This was however not the case. When measured by the Stanford Sleepiness Scale, no difference was found between the wake and sleep groups in any of the three test sessions. The two visual analogue scales also failed to find a difference between the two groups in the delayed and in the one-week follow up tests. These two latter scales did find a difference in the immediate test, with wake participants reporting better alertness, however this did not result in statistically significant effects in any of the tasks, and would not lead to the prediction that the more alert wake group ought to deteriorate more during the delay between the immediate and delayed tests.

An even stronger argument against circadian effects can be mounted by comparing performance in the delayed and the one-week follow up tests. Recall that these two tests took place at the same circadian time, with the sleep group performing both test sessions in the morning, and the wake group performing both sessions in the evening. Hence, if poorer performance in the wake group was due to time of testing, the same effect should have applied in the one-week follow up test. This was not the case however, with the wake group improving in all tasks from the delayed test to the one-week follow up, despite these two sessions both taking place in the evening. It appears then that circadian or fatigue effects are an unlikely source for the effects seen across these tasks.

To sum up the behavioural data, Experiment 8 succeeded in finding evidence for a sleep-related consolidation effect in a range of different word learning tasks. Both free recall and cued recall showed improving recall overnight, while a wake group either showed no change, or showed a decline over the course of a day. Importantly, in both tasks the wake group improved significantly when tested about a week later, at the same circadian time as the delayed test. This improvement was likely to be due to occurrence of sleep between the delayed and follow up tests. The same pattern was seen in RT gains in the old/new categorisation task. As expected, the lexical competition effect was not seen immediately after training, but did emerge after a night of sleep. Seeing this effect also in the wake group was an unexpected and novel finding to which I will return after considering the polysomnographic data.

While the behavioural data imply an important role for sleep as opposed to wake in consolidation of novel words, they do not on their own justify an active role for sleep (Ellenbogen, Payne, & Stickgold, 2006). For example, Wixted (2004, 2005) has argued that sleep plays a permissive role in memory consolidation by providing an environment free of interference. More specifically, he suggested that NREM sleep (like alcohol and certain drugs) blocks the induction of hippocampal long term potentiation (LTP), without disrupting the maintenance of previously initiated LTP, thus allowing consolidation to occur without interference. This explanation is in agreement with the sleep advantage seen in the above data. Interference can also be external in nature, for example the lack of linguistic input during sleep may allow the newly learned words to consolidate. An active role for sleep in memory consolidation on the other hand would be supported by identifying physiological events during sleep which are associated with the performance gains overnight, and would imply direct involvement in the consolidation process (Ellenbogen et al., 2006).

With relation to word learning, Experiment 8 sought to establish what these physiological events are by examining sleep stages and sleep spindles, both of which previous research has found to be correlated with memory consolidation. Looking at sleep stages first, SWS has previously been associated with consolidation of declarative memory, while REM sleep has been associated with procedural memory (although as discussed in the introduction, this dichotomy is likely to be too simplistic). Based on these observations, it was expected that measures of explicit word recall and recognition speed might benefit from SWS more than REM, and hence SWS duration may have predicted performance improvement overnight. No correlations between SWS duration (or any other sleep stage) and free recall and cued recall measures were found however. SWS duration did on the other hand predict overnight improvement in the old/new categorisation RTs, with longer SWS duration being associated with larger RT gains (although the statistical conclusions are weakened by multiple comparisons). As the old/new categorisation task requires an explicit decision to be made about the identity of the novel word, it can be classified as a declarative task and as such supports the dual process theory. One theory of why SWS might be particularly beneficial for consolidation of declarative memories was proposed by Gais and Born (2004b). According to this view acetylcholine modulates the direction of information flow between the hippocampus

and the neocortex. During wake and REM sleep, strong cholinergic activity suppresses feedback from the hippocampus to neocortex, but not feedback in the opposite direction. During SWS on the other hand, cholinergic suppression is released. Gais and Born (2004b) provided evidence for this position by chemically boosting cholinergic activity during SWS. This manipulation led to impaired consolidation of declarative memories overnight, but had no effect on consolidation of procedural memories. Also, the same manipulation during wakefulness had no effect.

Tononi and Cirelli (2006) have proposed an alternative theory of SWS function in memory consolidation. According to this view SWS (and magnitude of slow wave activity in particular) is crucial as it provides the neural environment for a decrease in synaptic connections. Learning during wakefulness results in an increase in synaptic strength and may result in saturated plasticity. SWS then provides an environment for rescaling of synaptic strength back to baseline levels. Thus the extent of learning and overnight change should be proportional to slow wave activity during the night. The data from Experiment 8 cannot arbitrate between the two views of SWS function, but support their shared view that SWS is important in declarative memory consolidation.

It remains to be explained why the SWS correlation was only seen in the old/new categorisation task, and not in the free and cued recall tasks. Firstly, it may be that there was not enough variability in these recall tasks for a correlation with SWS to emerge. As Figures 54 and 55 show, level of performance tended to be low in both tasks, and the changes overnight were numerically small. An RT measure on the other hand results in more variability, and may be better suited for a correlational design. Secondly, the variability in SWS duration across participants may also be too small for robust correlations. Gais and Born (2004b) argued that the lack of SWS correlations in many of the earlier studies suggests that the critical role of SWS is reliably observed only when large amounts of SWS are missing, such as in the split-night or SWS deprivation paradigms.

The second aspect of sleep physiology examined in this experiment was sleep spindle activity. No correlations were found between spindle activity and measures of novel word learning in free recall, cued recall, or the old/new categorisation task. Spindle activity did correlate with degree of lexical competition immediately after training, and change in lexical competition overnight. Participants who showed least

evidence for lexical competition (or a facilitatory effect where training of novel words resulted in faster RTs to their base words, i.e. the opposite of a competition effect) in the immediate test experienced higher levels of spindle activity during the following night. Also, increasing lexical competition effect overnight was associated with higher spindle activity. No correlations were found between spindle activity and lexical competition in the other test sessions.

To understand the association between spindle activity and lexical competition in the immediate test, it is important to note that many authors have argued that sleep benefits in memory consolidation are greater for weakly learned compared to strongly learned materials. In the declarative domain, Drosopoulos, Schulze, Fischer, and Born (2007) taught participants word-pairs either to a 90% correct or 60% correct criterion, and tested recall immediately after training and again after about 36 hours. The sleep group was allowed to sleep immediately after learning, while the wake group was sleep deprived during the night following training, but allowed to sleep on the night prior to the delayed test. While both groups showed a decline in recall rates, only the participants trained to the 60% criterion showed a sleep benefit (smaller decline) compared to their wake control group. No benefit of sleep was seen in the 90% criterion group, leading the authors to suggest that weakly encoded associations benefit from sleep more than strong associations. A similar conclusion was reached by Schmidt et al. (2006) who saw a correlation between spindle activity and word-pair recall only in a difficult condition using abstract words. Kuriyama, Stickgold, and Walker (2004) varied encoding difficulty in a procedural motor sequence task by varying sequence length and whether one or both hands were involved. While all sequences resulted in overnight improvement, the most difficult sequence involving both hands and nine elements showed the largest overnight improvement, again showing that poorly learned materials benefit most from overnight consolidation. It should be noted though that Tucker and Fishbein (2008) found sleep-dependent enhancement in declarative tasks only in participants who performed in the top half of a median split based on training performance. However, in this study sleep consisted of a nap with NREM sleep only making it difficult to evaluate whether the results were caused by the abnormal sleep or whether they would generalise to normal periods of sleep as well.

If it is accepted that sleep benefits weakly encoded memories more than strongly encoded memories, it is important to establish whether lack of lexical

competition (or presence of facilitation) reflects poor encoding. In the context of the current experiment, this can be done by correlating lexical competition in the immediate test with performance in the other tasks in the same test session. Unfortunately none of these correlations reached statistical significance, but they all showed a trend in the same direction. Less lexical competition was associated with poorer free recall ($r = 0.219$, $p = .09$), poorer cued recall ($r = 0.129$, $p = .33$), and slower old/new categorisation performance ($r = -0.061$, $p = .64$).

It is also possible to see theoretically how poor encoding of the novel words would result in facilitation of base word recognition (i.e., faster RTs to base words for which a new competitor has been trained). As the novel word is heard repeatedly during training, it is likely that it activates its phonologically overlapping base word (e.g., hearing *cathedruke* is likely to activate *cathedral*). Some evidence for this was seen in Experiment 1 where recall rates were higher in stimuli where the meaning suggested by the novel word form was related to the trained novel word meaning, showing access to base words upon learning the novel words. However, at the same time a neocortical trace for the novel word begins to gradually emerge, becoming stronger with each exposure. It is possible that the rate at which this new trace emerges varies between participants. For some participants the novel trace may become strong enough to begin to weakly compete with the base word already during training, reducing base word activation caused by training. In the immediate test these participants would show a small competition effect or a very small facilitation effect. For other participants the novel trace may not reach the necessary strength to begin to compete with overlapping base word representations, in these participants each novel word presentation during training would continue to activate the overlapping base word more strongly than the emerging new representation. Such repeated activation of the base word during training may result in base word facilitation in the immediate test.

If lexical competition (or the lack of it) in the immediate test is then seen as a measure of initial word learning, the association between base word facilitation in the immediate test and increased sleep spindle activity during the subsequent night support the notion that sleep is particularly beneficial for consolidation of memories that are weakly encoded during training. Recall also that there was a correlation between the amount of change overnight and spindle density, which remained marginally significant even when holding immediate competition constant. This also

supports the notion that sleep spindles play an important role in this consolidation process, and predict the degree of change in lexical competition overnight. Thus the present data may reflect two spindle-related processes of memory consolidation during sleep. Firstly, those participants who generated weak representations of the novel words seem to respond to the need to integrate the novel words in the lexicon by undergoing more spindle activity during the subsequent night. Secondly, the level of spindle activity appears to reflect the magnitude of consolidation overnight (measured in change in the competition effect), with higher activity resulting in larger change independent of the initial state of the novel word representations (although this partial correlation was statistically marginally significant). Such an interpretation is in line with theories of sleep spindle function, which suggest that spindles are a marker of hippocampal-neocortical information transfer, as described in the introduction.

Spindle activity was not equally distributed across the scalp. Frontal electrodes registered more spindles than the central electrodes. This is reminiscent of the data reported by Clemens et al. (2005), who found a correlation between verbal memory retention and spindle activity recorded at left frontal electrodes, and Schmidt et al. (2006) who reported a correlation between word-pair retention and spindle activity at frontocentral electrodes. In the current experiment left electrodes also registered more spindles than right electrodes. This latter finding may reflect the linguistic materials undergoing consolidation, although in the absence of a control condition this finding remains tentative. There was no clear disassociation between slow and fast spindles in the current experiment, with both contributing to the observed correlations. It should be noted that slow spindles are typically observed at frontal electrodes, while fast spindles dominate at parietal electrodes. This experiment did not include parietal recording sites, making it less likely to see a distinction between the two spindle types, as the central electrodes probably recorded a mixture of slow and fast spindles. Hence fast spindles may have been undercounted here. Future experiments should use a larger number of electrodes to better assess the contributions of the two spindle types in word learning.

One of the most interesting observations made in this experiment was that sleep spindle activity was associated only with emergence of lexical competition, and not with consolidation of explicit recall or recognition speed of novel words. It seems then that sleep globally enhanced novel word memory, possibly by

strengthening the neocortical trace, but sleep spindles were involved specifically in integrating novel words with overlapping existing words. To my knowledge this is the first demonstration of an association between spindles and a task which requires the integration of completely novel information with existing information. The CLS models predict that sleep should be most important in tasks such as this, which not only require strengthening of new memories, but also relating the new memories with existing memories, to allow generalisation of the newly acquired knowledge with respect to previously acquired experiences. These data also suggest that sleep plays an active role in the consolidation process, rather than just providing an interference-free environment.

Before leaving the spindle data, it is important to consider whether spindles might have a relationship with general learning ability. For example, Schabus et al. (2004) reported a marginally significant positive correlation between spindle activity and memory performance, leading them to suggest that spindle activity might be correlated with general learning aptitude. This was supported by a positive correlation between IQ and number of spindles reported by Nader and Smith (2003). In the present experiment spindle density correlated with improvement overnight (if improvement is measured by increasing lexical competition). Hence it would be possible to argue that the participants who improved the most may also have been the most gifted participants, or participants with the best learning capacity. However, if high spindle activity was a marker of better learning ability, it should have been possible to observe spindle correlations with all the word learning tasks, not just the task measuring lexical competition. Furthermore, spindle activity was also associated with low lexical competition in the immediate test rather than high competition which would have indicated good encoding of the novel words. These considerations suggest that the current data cannot by explained by a simple relationship between spindle activity and learning ability.

Finally, it is possible to reinterpret the behavioural lexical competition data in light of the sleep spindle findings. I proposed earlier some potential methodological reasons why the lexical competition effect emerged in the delayed test in the wake group, although data from Dumay and Gaskell (2007) suggested it should only be seen after sleep. The sleep architecture data suggested that even though the effect emerged also after a period of wake, there are physiological events during sleep (i.e., sleep spindles) which seem to be involved with lexical integration. This implies that

sleep may not be the only brain state in which lexical integration takes place, but it may be the optimal state, thanks to the unique properties of sleep physiology. This is not a novel position to take. Recently, Axmacher, Draguhn, Elger, and Fell (2009) have advanced a theory proposing that memory consolidation can take place during wake too. This is motivated, for example, by data showing that reactivation of brain areas related to new memories is seen not only during sleep but also during wake (Peigneux et al., 2006), and that hippocampal ripples occur at comparable rates during sleep and wake (Clemens et al., 2007). According to this integrative view, sleep still plays a unique role, for example in that it is involved in the synaptic downscaling proposed by Tononi and Cirelli (2006), but hippocampal-neocortical transfer may also occur during other resting states apart from sleep. Under this view the issue of how the wake group spends the delay between immediate and delayed test becomes vital. If many of the wake participants in the current experiment spent much of the delay in a resting state, this may have been enough to allow for some lexical integration to take place. Future studies should include a restful wakefulness control condition to fully disentangle the consolidation that takes place during sleep and potential consolidation during wake.

To summarise, Experiment 8 showed that sleep provides the optimal state for consolidation of meaningless novel words. This was seen as a sleep benefit in direct recall and recognition speed of novel words. SWS seems to be involved with consolidating this type of information, although here SWS duration correlated with the old/new categorisation task only. Lexical integration, as measured by novel words engaging in lexical competition with phonologically overlapping familiar words, was associated with sleep spindle activity. Low level of lexical competition in the immediate test was associated with higher spindle activity during the following night, and was proposed to reflect weak memory traces generated by novel word training. The magnitude of increase in lexical competition was also predicted by spindle activity, suggesting that sleep spindles are central in integrating newly learned memories with existing memories.

Chapter 7: Thesis summary and conclusions

The experiments reported in this thesis addressed two main questions. Firstly, to what extent is meaning useful or necessary in lexical integration (Chapter 2), and secondly, does the learning of novel word meanings benefit from offline consolidation to the same degree as learning of novel word forms appears to (Chapters 4 and 5)? Finally, an experiment elaborating on the role of sleep in offline consolidation of novel word forms was reported (Chapter 6). In the following section I shall summarise the main findings from each of these chapters.

7.1 Thesis summary

7.1.1 Chapter 2

Experiments 1-3 presented in Chapter 2 tackled the issue of whether meaning is necessary in generating new lexical representations that show evidence of having been integrated in the mental lexicon. As reviewed in Chapters 1 and 2, the previous literature regarding this question is mixed. Early work looking at access to new lexical representations in tasks such as letter identification and identity priming tended to show an advantage for meaningful novel words over meaningless novel words (Whittlesea & Cantwell, 1987; Balota et al, 1991; Rueckl & Olds, 1993; Rueckl & Dror, 1994). This conclusion however has not been consistently reached by studies looking at reading of novel words in children and adults (McKague et al., 2001; Nation et al., 2007; McKay et al., 2008), or by studies looking at novel object naming (James & Gauthier, 2004; Cornelissen et al., 2004; Gronholm et al., 2005, 2007). Two studies looking specifically at meaning and the integration of novel words in the mental lexicon also reached different conclusions about this issue. Dumay et al. (2004) found that both meaningful and meaningless spoken novel words engage in lexical competition to the same extent and with the same time course, while Leach and Samuel (2007) found that only meaningful novel words enable retuning of phoneme categories in a perceptual learning task.

In Chapter 2 I focused on the discrepancy between these two latter studies, and adopted the hypothesis that the different conclusions may have been affected by the nature of the novel words used in the two studies. Dumay et al. (2004) used novel words that overlap with existing words (e.g., *cathedruke*), while Leach and

Samuel (2007) did not. I referred to these two types of stimuli as neighbours and non-neighbours. Hence it is possible that hearing a neighbour novel word activates the meaning of its closest phonological real word neighbour, and that the neighbour novel word "inherits" a meaning in this way. To demonstrate that this is a plausible mechanism, in Experiment 1 I attempted to show that learners have access to the meaning of the real word from which the neighbour novel words were derived from. This was done by teaching participants meaningful neighbours and non-neighbours, where the meaning of the neighbour novel words was either consistent (e.g., *cathedruke* is a type of church) or inconsistent (e.g., *cathedruke* is a type of drink) with the meaning of the overlapping real word. This manipulation should only have an effect if the neighbour novel word evoked the meaning of the overlapping real word. Such an effect was seen already during training with better learning of meanings in consistent than in inconsistent neighbour novel words. The same was seen at test. Interestingly, at test a meaning consistency effect was seen even in cued recall of word forms. These data showed that learners have access to the meaning of the overlapping real words, and that this effect even extends to a task which does not explicitly require access to meaning (i.e., cued recall of word form).

The second half of the chapter tested the hypothesis that the nature of the stimuli was at least partially responsible for the discrepancy between the lexical competition and perceptual learning conclusions using both neighbour and non-neighbour novel word stimuli. Experiment 2 was a test of the ambiguous stimuli to be used in the novel word experiment, and demonstrated that these stimuli provide the standard perceptual learning effect in real words, whereby hearing the ambiguous phoneme /?dt/ in a word context (e.g., *awar[?dt]*) biased participants to categorise that ambiguous phoneme as a /d/ in a later phoneme categorisation test. This effect was seen immediately after training, suggesting that it does not require offline consolidation, and a day and a week later, suggesting that it does not benefit further from consolidation, but remains robust over several days.

Experiment 3 applied these ambiguous phonemes to novel words and showed that perceptual learning is not seen with non-neighbour novel words when no meaning is trained, consistent Leach and Samuel's (2007) conclusion. Crucially though, perceptual learning was seen with neighbour novel words, suggesting that some degree of meaning is necessary for lexical integration (at least when measured in perceptual learning) and that the inherited meaning in these stimuli is enough to

enable integration. Also consistent with Leach and Samuel, there was no evidence for a consolidation benefit in this experiment. I will return to this issue in a later section discussing consolidation in more detail.

7.1.2 Chapter 4

Having established that meaning plays a prominent role in the learning and lexical integration of novel words, in the next two experimental chapters I focused on the influence of offline consolidation both on learning novel word meanings and forms. As reviewed in Chapter 3, while a handful of studies have looked at semantic measures of lexical integration (Dagenbach et al., 1990; Perfetti et al., 2005; Breitenstein et al., 2007; Mesters-Misse et al., 2007, 2008; Dobel et al., in press), only one has directly assessed the role of offline consolidation in this process (Clay et al., 2007). I used semantic decision in Experiment 4 to measure speed of semantic access to novel words, half of which had been trained a day before testing and hence had had a chance to undergo consolidation for about 24 hours. While there was no speed advantage for consolidated over unconsolidated novel words overall, an RT advantage was seen in the last third of the task, providing preliminary evidence for a potential consolidation effect. This was interesting in light of explicit meaning recall data which showed better recall of unconsolidated words. A sentence plausibility judgement task supported the consolidation advantage in speeded access to meaning, although a later experiment suggested the effect in this task might have been orthographic rather than semantic. This experiment also included a shadowing task intended as a measure of access to phonological word forms, where a consolidation advantage was found in shadowing latencies and accuracy rates. As discussed in Chapter 5, the shadowing task is problematic however as it may include a non-trivial semantic component.

This shadowing effect was followed up in Experiment 5, using both shadowing and naming (reading aloud) tasks, and novel words that were meaningful or meaningless. The consolidation effect in shadowing was not replicated here, although in the naming task a consolidation effect was seen in error rates for meaningless novel words. Another task used to measure recall of novel word forms was cued recall, where a consolidation effect was found for both meaningful and meaningless novel words.

7.1.3 Chapter 5

The emerging consolidation effect in the semantic decision task in Experiment 4 indicated that offline consolidation may be a process of importance in learning novel word meanings. Experiments 6 and 7 put this theory to a stricter test by looking at semantic priming with novel word primes. While Breitenstein et al. (2007) showed cross-modal semantic priming using novel word primes, their paradigm used prime-target pairs that had been presented during training, and hence probably involved an episodic priming component. The experiments reported in Chapter 5 used a purer semantic priming paradigm where the test was to see if novel words prime not their own meanings but words associated with the novel word meanings. Furthermore, prime duration and SOA were manipulated in these experiments such that Experiment 6 measured priming with a large strategic component, while Experiment 7 was designed to tap into more automatic semantic activation. Both experiments showed that novel words can indeed prime associated real words, but only after a period of offline consolidation had been allowed to take place. Both priming types required consolidation, although the masked priming effect seemed to benefit from a longer consolidation period. In Experiment 6 with strategic priming the effect emerged with a 24 hour delay, and did not grow further significantly during the next six days. When looking at automatic priming in Experiment 7, reliable priming did not emerge immediately or 24 hours after training, but only when tested a week after training.

Again, explicit recall of meaning provided an interesting contrast to these priming data. In Experiment 6 recall was better on words learned on the day of testing compared with words learned one or seven days earlier. In Experiment 7, where only one set of words was trained and tracked over time, recall declined over one week, while priming emerged at the same time. Priming and explicit recall appeared to be dissociable, further strengthening the claim that priming measured a more automatic process of semantic activation.

Experiment 6 successfully replicated the shadowing effect, but also provided new information about the time course of this effect. Now no shadowing was found after a 24 hour consolidation opportunity (as in Experiment 5), but the effect did emerge if participants were given a one week consolidation opportunity. This

suggested that shadowing may benefit from an incremental, ongoing consolidation process over several days, much like the semantic priming effect.

7.1.4 Chapter 6

Experiment 8 looked at sleep-dependent consolidation in learning of spoken meaningless novel words. Dumay and Gaskell (2007) had shown that lexical competition effects require sleep to emerge. Experiment 8 examined sleep architecture during the post-training night, and sought to isolate the neural events during sleep that are associated with this type of lexical integration. The behavioural data showed sleep-associated performance improvements in free recall, cued recall, and novel word recognition speed overnight. Somewhat surprisingly the lexical competition effect emerged both after sleep and an equivalent time of wakefulness. Polysomnographic data showed that sleep spindle activity was associated with lexical integration. Participants who showed little evidence of competition immediately after training had higher spindle activity during the night following training. Spindle activity did not predict any other aspect of word learning. Time spent in slow wave sleep on the other hand was associated with magnitude of improvement overnight in the old/new categorisation task measuring novel word recognition speed. These data have significant consequences for our understanding of the cognitive role of sleep spindles. Spindles seem to be important in integrating new information with existing information, but less important in enhancing recall. I will discuss this finding further in Section 7.3 of this chapter.

7.2 Offline consolidation in word learning

One of the main aims of this thesis was to elucidate the time course of novel word learning, both in terms of learning the meaning of novel words and learning the form of novel words. Below I will compile the time course information derived from the different tasks used in the different experiments reported in the thesis, and look for consistent patterns across the experiments and related tasks.

7.2.1 Explicit recall of novel word meanings

All experiments in this thesis in which participants were taught meaningful novel words included a test of explicit meaning recall, where participants were asked to type in the meaning of each of the trained novel words. Table 9 summarises the outcome of these studies with regard to the contrast between consolidated and unconsolidated novel words. It also categorises experiments on whether an interference account might explain the difference between consolidated and unconsolidated conditions. As discussed in Chapter 5, this is a plausible account in those experiments where each participant learned two different sets of novel words before the test session (Experiments 4-6). Both proactive interference (PI) and retroactive interference (RI) have long been a focus of interest in research into forgetting (e.g., Underwood, 1945), with PI referring to the case where previous learning interferes with later learning, and RI to the case where later learning interferes with previous learning.

Table 9. Difference between explicit recall rates to novel word objects and features in consolidated (C) and unconsolidated (UC) conditions in each experiment.

	Lag	Objects	Features	Interference?
Experiment 1	Short	C = UC	C < UC	No
Experiment 4	Short	C < UC	n/a	Yes
Experiment 5	Short	C < UC	C < UC	Yes
Experiment 6	Short	C < UC	C < UC	Yes
	Long	C < UC	C < UC	Yes
Experiment 7	Short	C = UC	C = UC	No
	Long	C < UC	C < UC	No

Note: No features were trained in Experiment 4. Lag = Length of consolidation opportunity, Short = one day, Long = one week.

Table 9 shows that in the majority of experiments there was an advantage for unconsolidated, recently learned novel words compared to consolidated words which had been learned a day earlier. This pattern fits well with the notion of RI. However, most studies looking at RI have used a paradigm where participants first learn a list of cue-target pairs (A-B) and subsequently learn a new list using the same cues but different targets (A-C). Under these circumstances learning the second list impairs recall of the targets from the first list possibly as a result of competition between the

B and C targets (although other mechanisms are also possible, see Wixted, 2004, for a review). Although in the present experiments there was no reassignment of cues, RI is still a viable explanation for the data. Burns and Gold (1999) trained participants on a list of 120 familiar words, and tested recall immediately after the training. This was followed either by no further learning, learning of a new list of words using the same learning strategy as before, or the learning of a new list using a different learning strategy as before. When recall of the first list was subsequently retested, participants in the first two groups recalled more words than immediately after training, a consolidation-like phenomenon the authors referred to as "hypermnesia". Participants in the third group on the other hand recalled fewer words than in the first test, showing an RI effect. This condition was similar to the state of affairs in the experiments reported in this thesis, where a second set of novel words was learned using the same training regime as used in the first set of novel words. Hence, according to the view proposed by Burns and Gold (1999), the first set of novel words (the consolidated condition) would be subject to RI, and hence consolidation effects would be masked by interference-induced forgetting.

The potential interference effects observed in the current experiments may be seen as inconsistent with several reports of sleep-associated consolidation effects in declarative tasks. For example, Plihal and Born (1997) showed increased word-pair recall after a period of SWS sleep, and Lahl, Wispel, Willigens, and Pietrowsky (2008) showed increased recall of a list of words even after a brief nap compared to wake. Since presumably all participants in the present experiments slept between the training of the consolidated words and the test session, one might expect to see improved recall in this condition over the unconsolidated condition. However, there are two reasons why such effects might be masked in the experimental designs used in this thesis. The first reason is that sleep can have either an enhancing effect on memory, in which case we would see a consolidation advantage, or a protective effect (Ellenbogen et al., 2006, 2009) against decay. This latter effect would only be seen if sleep were directly compared with wakefulness over a period of time after learning. In such a comparison the prediction would be that the wake interval would result in a decline in recalled materials, while a sleep interval might be associated with no change or a smaller decline. As the present experiments measuring meaning recall did not involve such a contrast, it is not possible to fully relate the current findings with the existing sleep and consolidation literature.

The second reason why consolidation effects may be masked in these experiments has to do with the interference account. Table 1 shows that in all experiments where the interference account is plausible (i.e., experiments where two sets of novel words were trained), a significant advantage was seen for unconsolidated novel words, both in terms of recall of objects and number of features. In Experiment 1 where participants were trained on day 1 and recall was tested either immediately after training on day 1, or one day later on day 2, there was no significant difference between consolidated and unconsolidated conditions in recall of objects, although the difference between recall of features did reach significance. In Experiment 7 as well, novel words were trained on day 1 only. Here testing took place immediately after training, one day later, and again one week later. No difference was seen between day 1 and day 2 recall, but a significant decline had taken place by day 8. The contrast between day 1 and day 2 performance suggests that the declines seen in consolidated meanings in the other experiments may well have been at least partially caused by interference from learning a second novel word set (RI effect). Even so, no consolidation advantage was seen in the two experiments where the interference account can be eliminated. In Experiment 7 performance in the recall task was at ceiling, so this would have obscured any advantage. However, in Experiment 1 recall was clearly not at ceiling, hence the conditions there were favourable for a demonstration of a consolidation advantage. The fact that no such advantage was seen suggests that explicit recall of novel words does not benefit from enhancement over time. Experiment 7 further showed that recall declined over a week's time in the absence of interference, showing that there is no evidence in this task for a time-dependent consolidation process over a longer time scale either.

7.2.2 Speeded access to novel word meanings

While the meaning recall task measured accuracy of participants' explicit recall of novel word meanings, the experiments reported in this thesis also included tasks which were intended to measure speed of explicit and implicit access to novel word meanings. A summary of these data is shown in Table 10.

Table 10. Summary of findings in tasks measuring speed of access to meaning.

	Immediately after training	One day after training	One week after training
Semantic decision *(Experiment 4)*	- high accuracy	- accuracy decline - RTs faster	n/a
Strategic priming *(Experiment 6)*	- no priming	- priming found	- priming found
Automatic priming *(Experiment 7)*	- no priming	- no priming	- priming found
Sentence plausibility *(Experiment 6)*	- no evidence of change taking place over time in RTs or accuracy rates		

The semantic decision task in Experiment 4 showed declining response accuracy as a function of time of training, but revealed some evidence of gains in speed of access within a consolidation opportunity of one day. It should be noted though that this effect was only observed in the final third of the task, and only in the condition where the target and prime were related. The sentence plausibility task was conceptually closest to the semantic decision task, in that both tasks required a decision to be made about the congruency of the novel word meaning with provided context (a single word in semantic decision, a sentence in sentence plausibility judgement). The sentence plausibility task did not provide any evidence of consolidation effects when the task was modified such that the influence of possible form based processes was minimised (Experiment 4 vs. Experiment 6).

The priming experiments (Experiments 6 and 7) did not look at the speed of access to meaning directly, but rather looked at the influence the novel word meaning had on the processing of a real word presented shortly afterwards. By manipulating prime duration and SOA, I attempted to tap into semantic processes operating outside of strategic control in Experiment 7 and processes where strategic influences were available in Experiment 6. Experiment 6 (visible prime, long SOA) showed a priming effect only after consolidation had been given time to operate. The effect here was seen after a short consolidation opportunity of one day, and showed little evidence of growing further after a longer consolidation opportunity apart from being numerically smaller than the priming effect obtained with real word primes. This time course was consistent with semantic decision, which probably also benefits from explicit, strategic processes. Experiment 7 (masked prime, short SOA) also showed no priming immediately after training, but did reveal a priming effect of identical magnitude to the real word prime condition after a week of consolidation.

These experiments provide evidence that novel word meanings benefit from consolidation over time, and this process appears to be a gradual one operating over several days and/or nights.

As already discussed in Chapter 5, differences in the specific time course in the emergence of priming between the two types of priming may reflect a top-down boost in the case of semantic priming which includes a strong strategic component. When participants are aware of the primes and informed of the relationship between the primes and targets, they are more likely to try to actively access the meaning of the prime in order to facilitate lexical decision to the target. The semantic decision task of Experiment 4 suggested that such active access may speed up under certain circumstances over the first 24 hours after training. A primed lexical decision task, even with a long SOA, requires quick access to the meaning of the prime in order for priming effects to emerge. If the speed of such explicit access benefits from a short period of consolidation, then this gain would also translate into a priming effect in a task that encourages active use of the primes. Thus the consolidation gains seen in semantic decision and priming using a visible prime and long SOA may have their source in the same process, i.e. faster explicit access to meaning.

More automatic priming on the other hand would benefit less from faster explicit access. As discussed in Chapter 5, both multistage activation models and distributed network models rely on spreading activation between concepts to account for priming. In the absence of top down boost, spreading of activation can only occur when a new concept is sufficiently integrated in the semantic network, and/or when a new lexical representation is integrated with the semantic level. Experiment 7 suggests that such structural change takes more than one day or night to reach a state where the network can support masked priming. This view is consistent with the data reported by Clay et al. (2007), who found a semantic PWI effect only in their second test session which took place one week after training. Interestingly, they found a lexical PWI effect using novel words immediately after training (novel words interfering with picture naming irrespective of the semantic relationship between the word and picture), suggesting that a lexical representation may have been generated very quickly, but it had not been integrated with the semantic level until several days later. In sum, based on the present experiments, it seems that access to novel word meanings speeds up within a 24 hour period, which allows initial strategic priming effects to emerge. The necessary integration in the lexical-semantic system may

require a more fundamental change and takes more than one day or night to support effects detectable in masked priming. Interestingly the emergence of priming appears to be independent of participants' explicit recall accuracy of the novel word meanings. Unfortunately the sentence plausibility task seems to be too insensitive to pick up these effects, perhaps because the gradual presentation of the sentence introduces a delay between the relevant semantic processing and response execution which dilutes the effect.

7.2.3 Access to novel word forms

A number of tests across the experiments reported in this thesis also looked at recall of novel word forms. The tasks used included cued recall, free recall, shadowing speed and accuracy, reading speed and accuracy, and word recognition speed in old/new categorisation. The consolidation effects observed in these tasks are summarised in Table 11.

One of the prominent patterns seen in Table 11 is the data from the cued recall test. Here the two cued recall tests provided inconsistent results, with Experiment 1 showing no consolidation benefit in a group of participants tested one day after training compared to another group who were tested immediately after training. No difference between the groups was found. On the other hand, in Experiment 5 higher recall rates were seen in response to words learned one day before the test compared to words learned on the day of testing. While the experiments used different novel words, and the cue used in Experiment 5 was somewhat more difficult as it had more letters removed, overall success rates in the two experiments were similar, excluding potential ceiling or floor effects as possible explanations. Data from Experiment 8 provide a more detailed picture. In that experiment performance improvement was seen after a period of sleep, while an equivalent period of wakefulness resulted in declining recall rates. However, once both groups had had a chance to sleep, performance in the groups was equal (in the one-week follow up). This suggests that in this task change in performance is determined by the timing of sleep rather than simply time passing. Hence the lack of a consolidation effect in Experiment 1 can be explained by assuming that in the group which was tested a day after training recall rates initially declined during the day following the training session, and increased during the night prior to the testing

session, bringing recall rates back up to the original level. Experiment 8 also suggested that an overall improvement from immediate test could only have been seen if sleep had followed immediately from the initial training, protecting against the decline seen in the wake group.

Table 11. Summary of findings in tasks measuring access to word form knowledge in the consolidated (C) and unconsolidated (UC) conditions. Tasks showing evidence for consolidation are shaded.

	Time between UC and C conditions			
	10h delay	24h delay	1 week delay	Consolidation?
Cued recall *(Experiment 1)*		C = UC		No
Cued recall *(Experiment 5)*		**C > UC**		**Yes**
Cued recall *(Experiment 8)*	**Sleep: C > UC** Wake: C < UC		**Sleep: C > UC** Wake: C = UC	**Yes (sleep-dependent)**
Shadowing *(Experiment 4)*		**RT: C < UC** **Accuracy: C > UC**		**Yes**
Shadowing *(Experiment 5)*		RT: C = UC Accuracy: C = UC		No
Shadowing *(Experiment 6)*		RT: C = UC Accuracy: C = UC	**RT: C < UC** **Accuracy: C > UC**	**Yes (long delay only)**
Reading *(Experiment 5)*		RT: C = UC **Accuracy: C > UC ***		Yes (accuracy rates only)
Recognition speed *(Experiment 8)*	**Sleep: C < UC** Wake: C = UC		**Sleep: C < UC** **Wake: C < UC**	**Yes (sleep-dependent)**
Free recall *(Experiment 8)*	**Sleep: C > UC** Wake: C = UC		Sleep: C = UC Wake: C = UC	**Yes (sleep-dependent)**
Lexical competition *(Experiment 8)*	**Sleep: C > UC** **Wake: C > UC**		**Sleep: C > UC** **Wake: C > UC**	**Yes**
Perceptual learning *(Experiment 3)*		C = UC		No

Note: C refers to the consolidated condition, or when novel word performance is tracked over time to novel words in the consolidated state (i.e., 10h, 24h, or 1 week after training). UC refers to the unconsolidated condition or to words in the unconsolidated state (i.e., immediately after training). The signs "<" and ">" refer to differences in level of performance, which may be response time, recall rate, accuracy rate, or magnitude of the lexical competition or perceptual learning effect, depending on the task. * = in the meaningful condition only.

It is still left to be explained why a consolidation effect was seen in Experiment 5, an effect that would not be predicted by the above account. Recall that in Experiment 5 participants learned two different sets of novel words on two different days. Hence it is important to consider level of performance during training, as participants might have been more motivated to learn on one day than on the other. Cued recall tests carried out during the training sessions showed a small but statistically significant advantage for the session in which consolidated novel words were trained (day 1 training session). This may explain at least part of the consolidation benefit seen at test, although it should be noted that the difference between consolidated and unconsolidated words at test was larger than at training.

A similar time course may explain the lack of consolidation effects in the naming (reading aloud) task at least in terms of latencies. A consolidation benefit was seen in accuracy rates, but only in words for which a meaning was trained. The interpretation of the naming task is further complicated by the fact that naming appears to involve semantic processing. For example, semantic priming effects have been found in naming times to target words (see Neely, 1991, for a review), although the naming response does not require explicit semantic access. This means that the source of the consolidation effect in this task may be semantic, a view buttressed by the finding that the consolidation effect was restricted to meaningful novel words, and that access to novel word meanings can benefit from even a short consolidation opportunity of one day (Experiment 6).

A similar interpretation of the shadowing data was outlined in Chapter 5, where Experiment 6 revealed a consolidation effect in shadowing which seemed to operate over a longer time course than the shadowing data from Experiment 4 initially had implied. While Experiment 4 found a consolidation effect in shadowing in both latencies and accuracy rates after a short consolidation opportunity of one day, in Experiment 6 an effect was seen only after a long consolidation opportunity of one week. As discussed in Chapter 5, this suggested a long incremental consolidation process in this task, which may also be affected by semantic factors as all novel words in these two experiments were trained with meaning. Experiment 5 did not show a consolidation effect over one day, showing the effect to be fragile at a short time scale. This semantic hypothesis makes it more difficult to determine whether shadowing and reading tasks can be categorised primarily as form or

meaning based, as they are likely to reflect both processes. It must be left for future studies to further disambiguate the two influences.

Data from Experiment 8 can be treated with more confidence as no meaning was trained here. Both speed with which novel words are recognised (as measured in an old/new categorisation task) and free recall of novel word forms benefitted from consolidation over a night of sleep. In recognition speed improvement was seen over time, but the timing of sleep determined when the greatest gains were seen. In free recall a similar pattern was seen as in cued recall, with recall rates increasing during a night of sleep, with no change taking place during an equivalent time of wakefulness. However, semantic influences cannot be completely ruled out here either, as Experiments 1 and 3 showed that novel words overlapping with existing words may be influenced by the meaning of the neighbouring real words. This type of novel word was used in Experiment 8 in order to evaluate effects of lexical competition.

The lexical competition data were consistent with earlier reports of this effect emerging only after a delay (e.g., Gaskell & Dumay, 2003). The effect was not seen immediately after training, instead a significant facilitatory effect emerged whereby base words for which a new competitor had been trained showed faster RTs than base words for which no new competitor was trained. In contrast, about 10 hours later the competition effect was found, irrespective of whether the delay included sleep or wakefulness. This suggests that integrating novel words in the mental lexicon benefits from offline consolidation within a very short consolidation opportunity of up to 10 hours. The finding of this effect in both sleep and wake groups was unexpected in light of the data reported by Dumay and Gaskell (2007). Some potential explanations were discussed in Chapter 6, and the following section will present some more. In any case, some form of consolidation does appear to be crucial in this task.

While lexical competition is a measure of the degree to which a novel word has been integrated with neighbouring words at the lexical level, re-tuning of phoneme boundaries in the perceptual learning paradigm measures the degree to which novel word representations are able to affect a sublexical, phonemic level. Leach and Samuel (2007) found this effect in novel words immediately after training (provided that meaning had been trained too). The same conclusion was reached in Experiment 3, where a perceptual learning effect was found immediately after

training. This was also tested on the following day, however as the second test was preceded by a second training session, it is impossible to judge whether the effect had changed from day 1. There was however no effect of time, suggesting that at least the effect had not gained in strength overnight. Motivated by the absence of consolidation effects in the Leach and Samuel (2007) data, and data reported by Snoeren et al. (2009), Davis and Gaskell (2009) suggested that in the CLS framework the fast learning hippocampal system has a direct link to lexical phonology, negating the need for hippocampal-neocortical transfer to take place before phonology-related effects in novel word learning are observed. The data from Experiment 3 support this view.

7.3 Sleep in word learning

While most experiments in this thesis looked at the role of offline consolidation as a function of time including both sleep and wakefulness, Experiment 8 focused on sleep-specific consolidation. As reviewed in Chapter 3, this was motivated by a number of studies showing that sleep benefits various levels of language learning, and by the findings of Dumay and Gaskell (2007) who showed that lexical competition in novel words emerges after a night of sleep.

The primary finding in Experiment 8 was that sleep spindles seem to be involved in lexical integration, at least when measured by lexical competition. Spindle activity on the post-training night correlated with the magnitude of the lexical competition effect immediately after training, and the change in the effect overnight. As discussed in Chapter 6, if the lack of lexical competition or the presence of a facilitatory effect in the immediate test is taken as a sign of a weak lexical representation, then it appears that participants who generated weak novel lexical representations during training experienced more spindle activity overnight, suggesting that spindles are important in integrating novel words in the lexicon. This view fits well into the framework recently proposed by Stickgold (2009). According to this view the different sleep stages described in Chapter 6 not only correspond to consolidation of different types of memory (such as declarative and procedural), but may rather correspond to different consolidation processes. Stickgold (2009) proposed that SWS stabilises both declarative and procedural memories at a synaptic level, leading to recall enhancement. Stage 2 sleep and REM on the other hand serve

to facilitate systems level consolidation, which includes the hippocampal-neocortical transfer proposed by CLS models, extraction of rules and regularities from the newly learned information, and the integration of the new memories with existing memories. The data reported in Experiment 8 support this view. Of the tasks used in that experiment only the lexical competition measure indexes integration of novel words in the lexicon, and it was only this measure that showed an association with sleep spindle activity, one of the hallmarks of stage 2 activity. SWS duration on the other hand only correlated with improvements in novel word recognition speed, demonstrating the kind of dissociation predicted by Stickgold's (2009) account.

Stickgold (2009) also commented on the issue of whether weakly encoded memories benefit from sleep more than strongly encoded memories, by pointing out that while the majority of available data suggest that this is the case, some studies have found the opposite, and that this might mean that moderately well encoded memories benefit the most. This is because the brain might choose those memories for consolidation that benefit from the process the most. Memories that are already strongly encoded, and memories that are extremely poorly encoded are less likely to do so, as in the case of the former consolidation may be superfluous and in the case of the latter it may not be sufficient to retain the memory anyway. It is difficult to evaluate this hypothesis based on Experiment 8, as the data probably did not include a full range of encoding success levels, but it fits the idea that less strongly encoded memories benefit more from consolidation than strongly encoded ones.

As far as learning novel word forms is concerned, Experiment 8 suggested that integration in the lexicon benefits from sleep spindle activity, while enhanced recognition speed was associated with SWS duration. The explicit recall measures did not correlate with sleep stages or spindles. The role of sleep in the learning of novel word meanings on the other hand is unclear, and although the experiments reported in this thesis suggest meaning benefits from offline consolidation in general, they did not address the issue of sleep-specific consolidation. There is however good reason to believe that sleep plays an important role in the acquisition of meaning, and may in particular have been important in the semantic priming experiments. Stickgold, Scott, Rittenhouse, and Hobson (1999) tested semantic priming of weakly and strongly related word pairs during the day, and immediately after awakenings from REM and NREM sleep during the night. Larger priming effects were found in the strong priming conditions during the day and after

awakenings from NREM. However, after awakenings from REM the opposite was seen: now a larger priming effect was found in the weak priming condition than in the strong priming condition. These data suggest that REM sleep may play an active role in strengthening weak semantic associations, and hence may be of importance in learning novel word meanings and integrating the meanings in the semantic network. This hypothesis is supported by the current data where priming with novel word primes was only seen after a period of consolidation. The experiments reported in this thesis did not look at priming and sleep, but it is possible that if polysomnographic data had been collected, a correlation may have been found between the magnitude of novel word priming and REM duration on post-training night.

One study which did look at sleep and learning of semantic information found that sleep benefitted some aspects of semantic memory (Rogers & Mayberry, in preparation). In this study participants learned to name, recognise, and categorise satellites which could be typical (share several features with a category prototype) or atypical (share few features with a category prototype) exemplars of their category. Recall was tested after intervals including sleep or wakefulness. Sleep improved recall of atypical satellites more than typical satellites when the recall probe included both prototypical and individuating parts of the satellite. When the probe contained information only about individuating properties, the opposite pattern was seen with sleep improving memory of typical items more. The interaction is difficult to interpret, but at a minimum it seems that learning the satellite information benefitted from sleep to some degree, although the nature of the recall task seems to determine whether benefits for typical or atypical items are revealed.

Most sleep studies cited so far in this thesis show that one night of sleep following learning results in significant performance improvement. In contrast, the masked priming effect in Experiment 7 did not emerge after one night of sleep (or one day of wakefulness). This however should not be taken as evidence against a possible role for sleep in consolidation of meaning. It may be the case that consolidation continues over several nights after learning. While few studies have looked at this process in detail across several subsequent nights, there are reports from the procedural domain showing further performance improvements after the first night of sleep. Stickgold, James, and Hobson (2000) tracked visual discrimination performance for seven days after training and showed gradually

enhancing performance on the first four days (interestingly, no improvement was seen in participants who were sleep deprived for the first night after training). Both speed and accuracy improve more if measured after three nights of sleep compared to one night of sleep in a finger-tapping task (Walker, Brakefield, Seidman, Morgan, Hobson, & Stickgold, 2003). To my knowledge there are no similar data available from declarative tasks, but the masked priming (and shadowing) data reported in this thesis together with the procedural data cited above suggest that multiple nights of sleep may be important in meaning acquisition.

7.4 Limitations of the studies

The experiments in this thesis have been interpreted as dealing with L1 word learning, that is, participants are assumed to be treating the novel words as novel English words. It is however difficult to say how valid this assumption is. Participants might be treating the novel words as a class of information relevant only in the context of the experiment, especially as they know that the words and meanings are fictional. One way to try to avoid this state of affairs is to include training tasks where the novel words are presented in a naturalistic setting. In the current experiments looking at acquisition of meaning, this was attempted by including the sentence plausibility judgement task as part of the training. In this task participants were exposed to the novel words in English sentences, and asked to evaluate the appropriateness of the word in the sentence. This gave the participants at least some degree of experience with the novel words in an elaborate linguistic setting. It is difficult to estimate how successful these kinds of manipulations are though in the absence of much empirical data.

One of the few studies directly examining these issues was reported by Potts, St. John, and Kirson (1989). In these experiments participants read long stories that introduced new words and concepts (e.g., that *takahe* is a large flightless bird in New Zealand). One interesting manipulation concerned participants' beliefs about the veracity of the new information. Half of the participants were told that the information was correct, and the other half was told it was fictional. At test, which followed immediately after training, participants were asked questions about the meaning of the novel words either in a story context (defined by having a large number of filler questions directly related to the story) or a non-story context (with a

large number of fillers unrelated to the story). In the non-story context RTs to the questions were faster if participants thought the information was real compared to participants who thought it was fictional. The opposite was true in the story context. The authors argued that this suggests that only participants who thought the information was real encoded it in such a way that it was available in all contexts, not just the context of the experimental story.

Potts et al. (1989) also carried out a priming experiment (primed lexical decision) where the novel words acted as targets. Primes could be semantically related real words (concepts from the story), semantically unrelated but story-related (unrelated concept from the story) or completely unrelated (semantically unrelated and not occurring in the story). Story and non-story test contexts were again used. Relative to semantically unrelated primes (story-related or not), semantically related primes increased lexical decision accuracy rates in the story context only (RTs were not reported). No priming at all was found in the non-story context. This led the authors to argue that the new information was compartmentalised and not integrated with general world information.

The above data by Potts et al. (1989) suggest that experimentally trained novel words and their meanings may not be fully integrated in general world knowledge, at least when participants know that the words and meanings are fictional. If this was the case in the current experiments as well, this would undermine the assumption held in this thesis that the work here reflected normal L1 lexical processing. However, two considerations alleviate this concern. Firstly, the data presented in this thesis in fact seem to show that novel words were integrated with existing knowledge, both at the level of word forms (lexical competition) and at the level of novel word meaning (semantic priming). If the novel words had remained compartmentalised, these effects should not have been observed either immediately following training or after a delay. Secondly, Potts et al. (1989) tested participants only immediately after training. In light of the priming experiments reported here, it is possible that their priming and recall data might have benefitted from a period of offline consolidation prior to testing, possibly eradicating the compartmentalisation effect.

A related possibility is that participants may have thought of the novel words as words from a foreign language. This scenario is more difficult to exclude on the basis of the present data. Lexical competition effects can be seen between languages.

For example, Spivey and Marian (1999) in one of the early eye tracking studies showed that bilingual speakers of English and Russian when hearing an English target word (e.g., *bunny*) briefly also looked at a distracter object whose Russian name overlapped with the target (e.g., *bunka*, meaning jar), suggesting that words from the two languages engaged in lexical competition with each other. The same cross-language effect can be seen in semantic priming, where a prime in one language can facilitate responses to a related target in another language (e.g., Perea et al., 2008). The main strategy to try to get participants to think of the novel words as English words was to give them meanings in Experiments 5-7 for which no familiar word exists. Whether this was successful remains for future work to establish. In any case, it is not currently clear what the differences between learning new words in L1 and L2 are, this remains another interesting avenue for future work.

7.5 Conclusions and future work

In the sections above I outlined the findings of the thesis for offline consolidation with regard to learning novel word forms and meanings. In this concluding section I will draw those data together to present a time course of novel word learning. Figure 62 shows a timeline of word learning based on the data presented in this thesis.

In the figure 0 h refers to the situation immediately after training. As long as sufficient training has been provided, at this point participants usually exhibit good explicit recall of novel word meanings, as shown by high accuracy rates in the meaning recall tasks. Cued recall of word forms however is lower (Experiments 1 and 5), particularly in the auditory modality (Experiment 8). A new lexical representation appears to have been generated, as suggested by successful re-tuning of phonemic categories by perceptual learning (Experiment 3). While these new lexical representations seem to be able to influence sublexical levels, the lack of lexical competition and semantic priming effects indicates that these representations do not at this stage have fully functional links with other lexical representations or semantic representations.

Lexical competition was observed in the next step, at about 10 hours after training in Experiment 8 (12 h time point in Figure 62). Dumay and Gaskell (2007) have argued that this effect can only be observed after a night of sleep, but

	Task performance	Cognitive changes
0h	Perceptual learning observed. Good explicit recall of meanings and word forms.	Given sufficient amount of training, a novel lexical representation is established. The new representation is able to influence the phonemic level to enable re-tuning of phoneme boundaries.
12h	Lexical competition observed. If sleep has intervened, enhanced free recall, cued recall, and recognition time of novel words are seen.	Novel words integrate in the mental lexicon connecting with neighbours. Sleep facilitates this (sleep spindles in particular) but may not be necessary, as predicted by CLS accounts. Enhanced recall also suggests strengthening representation.
24h	Strategic semantic priming and faster semantic decisions observed. Shadowing times may begin to speed up and accuracy rates may increase.	Speeded access to novel word meanings is facilitated by now, reflected in semantic decision times and in emergence of strategic semantic priming. This may be a consequence of increasing lexical-semantic connection strengths.
24h - week	Automatic semantic priming observed. Shadowing latencies and accuracy rates increase.	Masked priming suggests the lexical-semantic integration has advanced substantially. Shadowing may also benefit from this semantic integration.

Figure 62. Timeline of novel words becoming part of the mental lexicon, based on data reported in this thesis.

Experiment 8 showed an effect in the wake group as well. However, this effect was associated with sleep spindle activity, showing that sleep does seem to play an active role in this form of lexical integration. Experiment 8 also showed other

developments that occur overnight, including enhanced free recall, cued recall, and novel word recognition speed when measured in an old/new categorisation task. Together these data indicate that new lexical representations integrate with neighbouring representations overnight, and also gain in strength resulting in enhanced recall and recognition. This change can be interpreted in the CLS accounts as a result of hippocampal-neocortical transfer and hippocampally driven reinstatement.

First novel word semantic priming effects emerge one day after training (24 h time point in Figure 62). This is the case for priming that includes a strong strategic contribution (long SOA, visible primes). It is important to note here that these effects too may be facilitated by sleep, the hypothesis about sleep-specific consolidation was not tested here. At this time point we can also see faster semantic decision times, suggesting that these developments may be caused by faster access to novel word meanings. This is possibly due to strengthening of lexical-semantic connections, or a reinstatement process strengthening a neocortical memory trace of the novel word meaning. Shadowing latencies begin to speed up and accuracy rates begin to improve at this time as well, potentially driven by the semantic changes. This particular finding needs to be interpreted carefully though, as Experiment 5 failed to see the effect at this time point.

At the final time point, beyond the change that occurs within 24 hours, we saw the emergence of semantic priming that relies more on automatic semantic activation rather than strategic access. Shadowing also seems to grow stronger over this longer time course. The exact time when these effects emerge is unclear, as in none of the experiments was there a testing time between 24 hours and seven days but it is clear that more than one day or night is required. However, as strategic top down influences are less likely to be helpful in this task, it can be hypothesised that the increased consolidation duration may be necessary for full integration of the newly learned words in semantic networks.

7.5.1 Main contributions of this thesis

The work reported in this thesis makes several contributions to the current understanding of novel word learning. These contributions can be summarised in four points.

1. Novel words that overlap with existing words (e.g., *cathedruke*) carry some degree of meaning inherited from the closest neighbour, and this meaning enables lexical integration when measured by re-tuning of phoneme categories, a finding not obtained with non-neighbours. This suggests that meaning is important in novel words becoming a fully functional part of the lexicon.

2. While explicit recall of novel word meanings does not seem to benefit from offline consolidation, speeded access to meaning does. Access to novel word meanings starts to become faster during 24 hours following training, enabling an initial strategic semantic priming effect to emerge with novel word primes one day after training.

3. Automatic spreading of semantic activation from novel words to existing words may not occur until more than one day or night of consolidation has been completed, reflected in emergence of masked semantic priming.

4. Sleep spindles are associated with integrating novel spoken word forms in the mental lexicon, but do not seem to be equally important in enhancing explicit memory or speed of recognition of the new words.

7.5.2 Future work

The conclusions outlined above raise a number of questions for future study. Here I will discuss some of the most important ones. Both the work presented in Chapter 2 and the experiments reported by Leach and Samuel (2007) suggest that meaning is important in lexical integration when measured by perceptual learning. However, evidence for this should be provided in other tasks of lexical integration. Most notably, it would be reassuring to see the same effect when lexical competition is used to index integration. This might be challenging as lexical competition studies require the use of neighbour novel words which I have argued carry meaning due to their overlap with existing words. However, it should be possible to teach participants new meaningless non-neighbours, and in a later training session introduce new meaningless and meaningful words that overlap with the previously

trained non-neighbours. Experiments along these lines could manipulate meaning in lexical competition.

As already mentioned in this chapter, one of the major questions concerns the role sleep plays in consolidation of novel word meanings. This could be examined in a number of ways. For example, the question of whether the strategic priming effect would be observed after a night of sleep (or even a brief nap) or an equivalent time of wakefulness would be easily resolved and might reveal a crucial role for sleep. The slower emergence of the masked priming effect could also be examined in a sleep study. The influence of learning on sleep architecture could be evaluated by polysomnograhically monitoring sleep both before training to obtain a baseline, and for several nights after training. Learning should have an impact on sleep, for example in the form of increased spindle activity. The number of nights it takes for this or other aspects of sleep architecture to return to baseline could be taken as a measure of consolidation duration. Ideally this return to baseline would coincide with the emergence of the priming effect.

Finally, I have argued in Chapter 6 that sleep spindles were closely associated with the integration of novel words in the mental lexicon, but less important in enhancing the explicit memory of the words. This is a potentially important clue about the role of spindles in memory consolidation in general. This finding should be expanded beyond the word learning paradigm to see if spindles have a broader function in relating new memories with existing memories.

Appendices

Appendix 1

Base words, neighbour novel words, non-neighbour novel words, and the meanings related to the base words used in Experiment 1.

Base word	***Neighbour***	***Non-neighbour***	***Meaning***
alcohol	alcohin	amcohin	*drink made of milk and has no taste*
amulet	amulos	abulos	*jewellery for priests and worn around the ankle*
assassin	assassool	illassool	*soldier who drives tanks and defends others*
baboon	babeel	baweel	*ape that lives in captivity and eats meat*
badminton	badmintet	rebmintet	*sport played outdoors with a heavy ball*
bayonet	bayoniss	deyoniss	*weapon made of wood and used in martial arts*
biscuit	biscan	liscan	*snack made of bamboo and tastes like corn*
bramble	bramboof	bromboof	*plant used a medicine and grows on mountains*
canyon	canyel	besyel	*valley that is shallow and has mud at the bottom*
caravan	caravoth	saravoth	*vehicle that runs on electricity and is slow*
cardigan	cardigile	mordigile	*clothing that has short sleeves and is made of wool*
cathedral	cathedruke	nathedruke	*church with metal benches and no windows*
clarinet	clarinern	clorinern	*instrument that is made of plastic and is shrill*
crocodile	crocodin	glacodin	*reptile that lives underground and eats roots*
daffodil	daffadat	saffadat	*flower that blooms in winter and is black*
dolphin	dolpheg	colpheg	*fish that is white and can't swim in deep waters*
dungeon	dungeill	mungeill	*prison for fraudsters built on an island*
fountain	fountel	cayntel	*spring found in the Arctic and produces sparkling water*
gimmick	gimmon	hummon	*trick done by amateur magicians and is learned quickly*
hormone	hormike	darmike	*chemical released in the cat liver to help digestion*
hurricane	hurricarth	pirricarth	*storm in tropical areas that lasts for months*
lantern	lantobe	lartobe	*torch that has a red light and is used by police*
mandarin	mandarook	hundarook	*fruit that grows in Siberia and is eaten by reindeer*
methanol	methanat	lethanat	*liquid that kills germs and is safe for humans*
molecule	molekyen	silekyen	*particle that is found in space and dangerous to manipulate*
napkin	napkess	dopkess	*cloth used to dry spilled drinks and is used in pubs*
octopus	octopoth	ardopoth	*animal that lives at the bottom of the sea and has small feet*
ornament	ornameast	ilnameast	*decoration made of dry leaves and used at Christmas*
parsnip	parsneg	corsneg	*vegetable with hard skin and tastes sour*
pelican	pelikibe	nelikibe	*bird found in cold areas and builds a nest on ice*
pyramid	pyramon	dyramon	*building made of marble and used to store books*
skeleton	skeletobe	speletobe	*bone in the knee and dislocates easily*
spasm	spasel	trasel	*cramp one gets after swimming and happens during sleep*
squirrel	squirrome	speirrome	*rodent that is hairless and has no ears*
tavern	tavite	tapite	*bar for vegetarians and serves expensive food*
yoghourt	yogem	yegem	*dessert made of butter and fresh mint*

Appendix 2

Word stimuli used in Experiment 2.

/d/-ending words	/t/-ending words
aloud	acquit
amend	acute
amid	amulet
arcade	carrot
award	chocolate
collide	emit
commode	equate
comrade	hermit
corrode	ignite
coward	inert
elude	innate
erode	locate
grenade	merit
horrid	negate
lemonade	nominate
liquid	ornate
maraud	recruit
marmalade	regret
orchard	unite

Appendix 3

Neighbour novel words and their corresponding derived non-neighbour novel words, used in Experiment 3.

Neighbours		Non-neighbours	
/t/-ending words	/d/-ending words	/t/-ending words	/d/-ending words
alcohoite	alcohoide	umbohoite	umbohoide
babort	babord	nenort	nenord
bayonout	bayonoud	royenout	royenoud
canyit	canyid	gemyit	gemyid
gimmort	gimmord	laggort	laggord
hormart	hormard	thernart	thernard
hurriceet	hurriceed	bolliceet	bolliceed
methanat	methanad	poranat	poranad
molekyet	molekyed	karekyet	karekyed
napket	napked	rebket	rebked
ornamert	ornamerd	arlimert	arlimerd
peliket	peliked	mabiket	mabiked

Appendix 4

Novel words used in Experiments 4-7. Words used in Experiment 7 are marked with an asterisk.

agglem*	horand	terum
aggrood	horrot	tobbin*
anniby	jabbary*	tobir
ardoff*	jeprium	uvar*
arifie	jommer	valliss
blontack*	kerple*	velchur
bochor	konrith*	vilchy*
boumnet	lanbir	virrin
chebbor*	lerret	volbor
chisdow*	liddim	vorent*
cosmer	lidgy	vuckor
criddin	limmout	vurrith
dawtatt*	liutist	waba*
daxon	loodit*	whadal
distap	luddilat	whummith
dobbir*	lupitat*	wiblid
dunnath	mearton	
elch	meckalen*	
entelem*	merdut*	
eritriff*	milgium	
erotron	mippun	
fammar	molbit	
feckton*	onnith	
feffsol	ospont*	
femmet	pannetor	
fevous	peckolet*	
fiddioth	poffren*	
flimmir*	ponniol	
foostel	quammish*	
frocant	quellit	
gahoon*	quellop	
gettent	quemmer*	
gittow	siffor	
glain*	slethy*	
goitrem	sluckmor	
gomth	smetton	
goolat	sockadol	
gordar	somture	
gworp	soppimin	
hentun	speth*	
heprit*	spuffron	
hoddar*	sudoid	
hoddit	teffien	

Appendix 5

Novel word meanings, associated target words, and target nonwords derived from the words, used in Experiment 4. Same targets but elaborated meanings (Appendix 8) were used in Experiment 6. The subset of stimuli used in Experiment 7 is marked with an asterisk (but see Appendix 8 for elaborated meanings). The three targets are ordered in descending order of association strength.

Meaning	Target 1	Target 2	Target 3	Nonword 1	Nonword 2	Nonword 3
baby*	child	cry	infant	chyld	cro	inlant
bath	shower	soap	tub	shoger	woap	tur
battery	car	acid	charge	cas	ocid	cherge
beef*	steak	meat	roast	steat	veat	poast
bench	sit	seat	chair	sut	seak	cheir
blanket	warm	cover	sheet	garm	coser	sheeg
boat	motor	sail	ship	motot	cail	shup
bone	skeleton	break	marrow	skemeton	breal	varrow
book*	read	school	study	pead	schood	stidy
bread*	butter	dough	loaf	vutter	mough	loat
candle*	light	wax	flame	jight	wex	flome
cat*	dog	mouse	kitten	dox	wouse	kitgen
chicken	soup	wings	bird	soug	hings	birv
closet	clothes	door	hanger	cluthes	doot	henger
cloud	sky	rain	white	sby	rait	whote
coal	black	mine	fuel	blyck	gine	fuer
coat	jacket	hat	cold	jacken	har	nold
cow*	milk	calf	bull	rilk	calt	jull
cream*	whip	coffee	cheese	whis	coftee	cheete
crowd	people	mob	group	peaple	wob	groud
crown*	king	jewel	queen	fing	jefel	queel
desert	sand	dry	hot	nand	dro	fot
desk	lamp	table	work	lamf	mable	mork
drawing*	art	picture	sketch	ast	pictere	skitch
ear*	hear	sound	head	vear	soind	heax
face*	eyes	nose	smile	oyes	nosa	smige
farm	crops	country	barn	frops	coustry	barg
fist*	fight	hand	punch	feght	hond	ponch
fog*	mist	smog	thick	misp	swog	theck
gate	fence	open	entrance	wence	opet	entrynce
guitar	string	music	piano	streng	rusic	pieno
hill	mountain	climb	slope	moubtain	dlimb	slore
horse	ride	pony	saddle	rige	pona	laddle
hospital	ill	nurse	bed	ild	numse	bep

Continued on the next page

Appendix 5 continued

Meaning	Target 1	Target 2	Target 3	Nonword 1	Nonword 2	Nonword 3
knee	ankle	bend	joint	ankla	jend	joing
knife*	fork	cut	blade	fosk	vut	blada
knight*	armor	soldier	sword	arhor	solpier	swort
leg*	arm	body	walk	arn	sody	walp
lemon*	lime	sour	orange	limi	rour	orenge
lid	cap	top	jar	cip	tep	sar
lock	key	close	secure	koy	clase	secufe
maid*	clean	servant	butler	cleah	sermant	bunler
map	world	direction	travel	worlt	diriction	trovel
meadow*	field	grass	flower	fielm	prass	flowen
mirror*	reflection	image	glass	seflection	umage	gless
missile	war	rocket	bomb	wir	rockel	gomb
monk*	priest	monastery	religion	proest	modastery	teligion
moon	sun	star	night	lun	stur	vight
neck*	shoulder	throat	tie	shounder	throad	kie
needle*	thread	sew	pin	threal	gew	pid
ocean	sea	water	wave	nea	jater	wahe
pan*	pot	cook	fry	pog	wook	bry
path*	road	trail	way	roat	truil	woy
pill*	medicine	drug	aspirin	tedicine	drig	asmirin
pistol*	gun	shoot	rifle	gug	shoon	rikle
plate	dish	food	eat	desh	foor	eam
prison*	jail	bar	cell	jais	ber	rell
radio	television	stereo	station	velevision	stebeo	statien
ring*	finger	wedding	diamond	cinger	wodding	diawond
seed	plant	sow	grow	plent	fow	grom
sheep*	wool	lamb	herd	woot	pamb	hird
shoe*	foot	sock	lace	foet	seck	labe
skirt*	dress	blouse	shirt	driss	blousa	shirf
telephone	call	number	talk	rall	numbem	tald
tent	camp	hut	shelter	mamp	huv	shebter
tooth*	decay	ache	brush	debay	uche	bresh
tractor	machine	dirt	pull	vachine	dirf	tull
tree	leaf	trunk	stump	keaf	trynk	stumb

Appendix 6

Real word primes, associated targets, and nonword targets derived from the word targets used in Experiments 4-7. The three targets are ordered in descending order of association strength.

Prime	Target 1	Target 2	Target 3	Nonword 1	Nonword 2	Nonword 3
ambulance	emergency	siren	accident	emermency	giren	accicent
balloon	air	helium	float	oir	hesium	fload
binder	folder	notebook	paper	volder	notegook	waper
bruise	hurt	pain	hit	hurp	pait	hib
burglar	thief	robber	steal	thiel	tobber	steab
cannon	ball	fire	weapon	byll	fite	weanon
circus	clown	animal	carnival	clewn	animad	carpival
clinic	doctor	sick	health	hoctor	bick	heamth
coffin	dead	burial	grave	sead	butial	frave
dart	board	game	throw	boarf	gamu	thriw
eraser	pencil	mistake	rubber	pencid	misvake	subber
flask	wine	bottle	whiskey	wina	bittle	whilkey
flour	cake	bake	sugar	cace	dake	sutar
frog	toad	hop	jump	toak	kop	fump
herb	spice	tea	garden	spoce	toa	larden
ketchup	mustard	red	tomato	muskard	rer	togato
lizard	reptile	snake	green	reppile	sneke	dreen
medal	gold	award	honor	golp	awarn	hosor
nun	convent	church	sister	lonvent	chorch	tister
oyster	clam	shell	pearl	claf	shull	pearn
paddle	row	oar	canoe	rop	oad	casoe
parcel	package	post	box	dackage	posk	jox
pebble	rock	stone	beach	vock	stine	neach
raisin	grape	prune	fruit	grare	prane	frait
salad	lettuce	dressing	bowl	lettace	bressing	bewl
sausage	breakfast	pork	bacon	breamfast	porl	gacon
slug	worm	slow	snail	worb	sfow	snoil
termite	bug	wood	pest	byg	bood	mest
tiger	lion	jungle	stripe	liot	dungle	strepe
towel	cloth	wet	wash	clath	det	wosh
vampire	blood	bat	fangs	bloot	baf	nangs
vinegar	oil	bitter	salt	oid	vitter	sall
wallet	money	purse	leather	momey	pursa	leathen
wasp	sting	bee	nest	stong	kee	nesk

Appendix 7

Sentences used in the sentence plausibility judgement task in the test session of Experiment 4.

Meaning	Sentence
baby	The parents were proud to announce the birth of the
bath	The girl loved relaxing in warm water so once a week she took a long
battery	The salesman boasted that the mobile phone was powered by a strong
beef	The man didn't care for vegetarian food so he chose a burger with
bench	The tired shopper decided to rest for a while on a wooden
blanket	The woman was cold so she wrapped herself in a wool
boat	The fisherman was sad after the sinking of his
bone	The nurse fitted the girl with a cast to allow the healing of the
book	The librarian could not find the
bread	The woman living next to a bakery loved the smell of fresh
candle	The man was mindful of fire safety and put out the
cat	The woman liked to listen to the purring of her
chicken	The farmer couldn't produce enough eggs because he had only one
closet	The businessman kept his suits neatly in his
cloud	The eclipse was covered by a large
coal	The boy searched the mine for gold but only found a lump of
coat	The girl was freezing so the gentleman offered her his
cow	The vet inspected the hooves of the
cream	The lovers fed each other strawberries and
crowd	The dictator feared the demonstration of the angry
crown	The prince hoped that one day he could carry on his head the
desert	The archaeologist found a pharaoh's tomb in the middle of the
desk	The office worker tried to work late but fell asleep on his
drawing	The parents were impressed when the child painted a lovely
ear	The doctor told the patient the loud music had damaged the drum of his
face	The man always wore a mask to hide his disfigured
farm	The head of the agriculture department had himself too grown up on a remote
fist	The man was furious and hit the table with his
fog	The plane could not land due to a heavy
gate	The guards saw the enemy approach and closed the castle's
guitar	The man sang a serenade to the girl while playing his
hill	The driver found that the car struggled to get up the
horse	The man stood by the track and cursed himself for betting on the wrong
hospital	The ambulance crew had only minutes left to get the patient to the
knee	The young boy sat happily on his grandfather's
knife	The cook sliced the vegetables with his
knight	The maiden locked in the tower was rescued by a handsome
leg	The athlete couldn't run after breaking his
lemon	The man preferred his iced tea with a fresh slice of
lid	The grandmother wanted to eat the jam but couldn't open the
lock	The guard could not stop the people from opening the door because of a broken
maid	The man didn't have time for housekeeping so he hired a professional
map	The tourist guide marked the location of the museum on the
meadow	The children ran out and rolled in the dewy
mirror	The girl enjoyed watching herself in the
missile	The submarine was carrying one nuclear
monk	The man enjoyed meditating so much that he became a Buddhist
moon	The astronaut landed on the
neck	The man found the shirt otherwise comfortable but the collar was too tight around his
needle	The tailor pricked his finger with the

Appendix 7 continued

Meaning	Sentence
ocean	The diver found the remains of the ship at the bottom of the
pan	The chef made an omelette on his non-stick
path	The hiker got lost after following the wrong
pill	The patient needed a glass of water to swallow the
pistol	The sheriff threatened the robbers with his
plate	The man was so hungry that he devoured everything on his
prison	The judge sentenced the criminal to two years in
radio	The grandfather never forgot to listen to the daily news on his
ring	The man asked her to marry him and gave her an expensive
seed	The man ate a piece of melon and spit out a black
sheep	The shepherd was horrified when he saw in the field only one
shoe	The woman broke one of her heels and needed to buy a new
skirt	The father objected to his daughter's short
telephone	The woman was in the shower when she heard the ringing and rushed to answer the
tent	The hunters chose a clearing in the forest and spent the night in their
tooth	The dentist pulled out the patient's
tractor	The farmer annoyed the motorists by driving on the road in his slow
tree	The boys competed in who was the fastest to climb to the top of the

Appendix 8

Elaborated novel word meanings used in Experiments 5 and 6 (object and two features). The subset of meanings used in Experiment 7 is marked with an asterisk.

Simple meaning	Elaborated meaning
baby*	type of baby that is premature and underweight
bath	type of bath that is marble and oval
battery	type of battery that is lightweight and environmentally friendly
beef*	type of beef that is British and comes from calves
bench	type of bench that has six sitting spaces and is metal
blanket	type of blanket that is a quilt and made of wool
boat	type of boat that is made of fibreglass and is the same size as a car
bone	type of bone that is flexible and part of a backbone
book*	type of book that has pictures and is oversize
bread*	type of bread that is dark brown and has nuts in it
candle*	type of candle that has a fragrance and has an especially bright flame
cat*	type of cat that has stripes and is bluish-gray
chicken	type of chicken that is mostly red and has feathery feet
closet	type of closet that has a curtain and is spacious inside
cloud	type of cloud that is purple and appears at sunset
coal	type of coal that is energy-efficient and used in barbeques
coat	type of coat that is waterproof and warm
cow*	type of cow that has a hairy tail and has giant horns
cream*	type of cream that is organic and low in fat
crowd	type of crowd that is angry and without a leader
crown*	type of crown that is worn by monarchs and is made of rubies
desert	type of desert that is in Western China and is expanding
desk	type of desk that has wheels and is made out of plastic
drawing*	type of drawing that is a portrait and is in neon colours
ear*	type of ear that belongs to a mammal and is folded
face*	type of face that has had plastic surgery and looks completely different
farm	type of farm that grows livestock and is located in South America
fist*	type of fist made with the thumb on top and a bent wrist
fog*	type of fog that happens in equatorial areas and appears very quickly
gate	type of gate that is steel and is alarmed
guitar	type of guitar that is made of mahogany and is expensive
hill	type of hill that has no grass on it, and is found in cold regions
horse	type of horse that races and has curly hair
hospital	type of hospital that treats patients with depression and is located in the U.K.
knee	type of knee that is bony and has been previously broken
knife*	type of knife that is often used by butchers and is very sharp
knight*	type of knight that carries a banner and protects the helpless
leg*	type of leg that that is long and very muscly
lemon*	type of lemon that is seedless and imported from Mexico
lid	type of lid that forms a seal and is heavy
lock	type of lock that is voice-controlled and is hard to break
maid*	type of maid that comes in once a day and takes care of pets
map	type of map that shows where treasure is buried and is made of parchment

Appendix 8 continued

Simple meaning	Elaborated meaning
meadow*	type of meadow that buffalo graze in and that was created by Native Americans
mirror*	type of mirror that is circular and is convex
missile	type of missile that is equipped with a nuclear explosive and is made in North Korea
monk*	type of monk that lives in Tibet and fasts for seven days at a time
moon	type of moon that is bright and crescent shaped
neck*	type of neck that is short and freckled
needle*	type of needle that is made of platinum and can make very small stitchs
ocean	type of ocean that is polluted and where the whale population is decreasing
pan*	type of pan that is battery-heated and used for camping
path*	type of path that is paved and occurs in parks
pill*	type of pill that lowers cholesterol and blood pressure
pistol*	type of pistol that carries 20 bullets and can fire very quickly
plate	type of plate that is square-shaped and made of wood
prison*	type of prison that is for murderers and is located in the U.S.
radio	type of radio that uses solar power and has a square aerial
ring*	type of ring that is silver and engraved
seed	type of seed that is dimpled and comes from tropical fruit
sheep*	type of sheep that lives in Scotland and has soft hair
shoe*	type of shoe that has a strap and is made of plastic
skirt*	type of skirt that is flowery and made of silk
telephone	type of telephone that can make video transmissions and is portable
tent	type of tent that is solar heated and used in remote places
tooth*	type of tooth that is weak and is discoloured
tractor	type of tractor that is used for carrying bulky loads and uses diesel fuel
tree	type of tree that lives for over 200 years old and is very tall

Appendix 9

Sentences used in the sentence plausibility judgement task in the test session of Experiments 5 and 6.

Meaning	Sentence
baby	The doctor was happy to announce the survival of the
bath	The girl loved relaxing in warm water so once a week she spent an hour in the
battery	The salesman boasted that the mobile phone was powered by the
beef	The man didn't care for vegetarian food so he chose a burger with
bench	The tired shopper decided to rest for a while on the
blanket	The woman was cold so she wrapped herself in the
boat	The sailor was sad after the sinking of his
bone	The nurse fitted the girl with a brace to allow the healing of the
book	The librarian could not find the
bread	The woman living next to a bakery loved the smell of fresh
candle	The man was mindful of fire safety and put out the
cat	The woman liked to listen to the purring of her
chicken	The farmer couldn't produce enough eggs because he had only one
closet	The businessman kept his suits neatly in his
cloud	The sky was covered by a large
coal	The boy searched the mine for gold but only found a lump of
coat	The girl was freezing so the gentleman offered her his
cow	The vet inspected the hooves of the
cream	The lovers fed each other trifle with
crowd	The politician feared the demonstration of the
crown	The princess hoped that one day she could carry on her head the
desert	The paleontologist found a dinosaur's bone in the middle of the
desk	The office worker tried to work late but fell asleep on his
drawing	The parents were impressed when the child painted a lovely
ear	The doctor told the old lady the loud music had damaged the drum of her cat's
face	The man felt confident for the first time because of his
farm	The head of the agriculture department had also grown up on a remote
fist	The man was furious and hit the table with his
fog	The plane could not land due to a heavy
gate	The guards saw the terrorists approach and closed the building's
guitar	The guitarist sang a serenade to the girl he loved while playing his
hill	The driver found that the car struggled to get up the
horse	The man stood by the track and cursed himself for betting on the wrong
hospital	The mental illness carers advised the patient to go to the
knee	The young boy sat happily on his grandfather's
knife	The cook sliced the lamb with his
knight	The maiden locked in the tower was rescued by a handsome
leg	The athlete couldn't run after breaking his
lemon	The man preferred his iced tea with a fresh slice of
lid	The grandmother wanted to eat the jam but couldn't open the
lock	The guard stopped the people from entering the room by activating the
maid	The man didn't have time to take care of his guinea pigs so he hired a professional
map	The museum exhibited the crumbling paper of a 15th century
meadow	The children ran out and rolled in the dewy
mirror	The girl enjoyed watching herself in the
missile	The submarine was carrying one
monk	The man enjoyed meditating so much that he became a deeply religious
moon	The poet wrote a poem describing the sky full of stars and a beautiful
neck	The man found the shirt otherwise comfortable but the collar was too tight around his

Appendix 9 continued

Meaning	**Sentence**
needle	The woman fixed a hole in her child's clothing with the
ocean	The diver found it difficult to see the remains of the ship at the bottom of the
pan	The wife made an omelette on her non-stick
path	The old man got lost after following the wrong
pill	The patient needed a glass of water to swallow the
pistol	The sheriff threatened the highwayman with his
plate	The man was so hungry that he devoured everything on his
prison	The judge sentenced the criminal to 100 years in the
radio	The grandfather never forgot to listen to the daily news on his
ring	The man asked her to marry him and gave her an expensive
seed	The man ate a piece of melon and swallowed a large
sheep	The owner of the farm was horrified when he saw in the field only one
shoe	The woman did not notice that the lace had become untied in her left
skirt	The father objected to his daughter's short
telephone	The woman was in the shower when she heard the ringing and rushed to answer the
tent	The hunters chose a clearing in the forest and spent the night in their
tooth	The dentist pulled out the patient's
tractor	The farmer annoyed the motorists by driving on the road in his slow
tree	The boys competed in who was the fastest to climb to the top of the

Appendix 10

Base words, novel words, and novel word foils used in Experiment 8.

Base word	Novel word	Foil
alcohol	alcohin	alcohid
amulet	amulos	amulok
artichoke	artiched	artichen
assassin	assassool	assassood
baboon	babeel	babeen
badminton	badmintel	badmintet
bayonet	bayoniss	bayonil
biscuit	biscal	biscan
blossom	blossail	blossain
bramble	brambooce	bramboof
canvas	canvick	canvit
canyon	canyel	canyes
capsule	capsyod	capsyoff
caravan	caravoth	caravol
cardigan	cardigite	cardigile
cartridge	cartroce	cartrole
cataract	catarist	catarill
cathedral	cathedruke	cathedruce
consensus	consensom	consensog
crocodile	crocodiss	crocodin
culprit	culpren	culpred
daffodil	daffadat	daffadan
decibel	decibit	decibice
dolphin	dolpheg	dolphess
dungeon	dungeill	dungeic
fountain	fountel	founted
gelatine	gelatord	gelatorl
gimmick	gimmon	gimmod
grimace	grimin	grimib
haddock	haddale	haddan

Base word	Novel word	Foil
hormone	hormike	hormice
hurricane	hurricarb	hurricarth
hyacinth	hyasel	hyased
lantern	lantobe	lantoke
mandarin	mandarook	mandarool
methanol	methanack	methanat
mistress	mistrool	mistrooke
molecule	molekyen	molekyek
moped	mopall	mopass
mucus	muckip	muckin
napkin	napkem	napkess
octopus	octopoth	octopol
onslaught	onsleete	onsleeth
parachute	parasheff	parashen
parsnip	parsneg	parsnes
partridge	partred	partren
pedestal	pedestoke	pedestode
pelican	pelikiyve	pelikibe
profile	profon	profod
pulpit	pulpen	pulpek
pyramid	pyramon	pyramotch
siren	siridge	sirit
skeleton	skeletobe	skeletope
slogan	slowgiss	slowgith
spasm	spaset	spasel
specimen	specimal	specimav
squirrel	squirrome	squirrope
tavern	tavite	tavile
tycoon	tycol	tycoff
utensil	utensont	utensop

Appendix 11

Correlations between the lexical competition effect and sleep spindle activity at each of the four electrodes.

		C3		C4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	**0.601**	0.332	**0.651**	0.321
	p	***0.002***	*0.11*	***> 0.001***	*0.10*
Immediate	*r*	**-0.554**	-0.396	**-0.622**	-0.370
	p	***0.005***†	*0.06*	***0.001***	*0.06*
Delayed	*r*	0.282	0.051	0.270	0.056
	p	*0.18*	*0.81*	*0.17*	*0.78*
Follow up	*r*	0.123	0.109	0.220	0.091
	p	*0.57*	*0.62*	*0.28*	*0.66*

		F3		F4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	**0.589**	0.335	**0.495**	0.200
	p	***0.001***	*0.09*	***0.012***†	*0.34*
Immediate	*r*	**-0.542**	-0.375	**-0.447**	-0.169
	p	***0.003***	*0.054*	***0.025***†	*0.42*
Delayed	*r*	0.211	0.035	0.284	0.125
	p	*0.29*	*0.86*	*0.17*	*0.55*
Follow up	*r*	0.082	0.084	0.133	0.135
	p	*0.69*	*0.68*	*0.53*	*0.53*

Note: Significant correlations in bold. Density = spindle density (number of spindles per 30 seconds), Ampl. = average maximal spindle amplitude.

Continued on the next page

Correlations between free recall of novel words and sleep spindle activity at each of the four electrodes.

		C3		C4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	-0.287	-0.182	-0.219	0.021
	p	*0.17*	*0.40*	*0.27*	*0.92*
Immediate	*r*	-0.394	0.159	-0.201	0.176
	p	*0.057*	*0.46*	*0.31*	*0.38*
Delayed	*r*	**-0.488**	0.001	-0.293	0.156
	p	***0.016***†	*0.99*	*0.14*	*0.44*
Follow up	*r*	**-0.475**	-0.027	-0.334	-0.023
	p	***0.022***†	*0.90*	*0.10*	*0.91*

		F3		F4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	-0.308	-0.05	-0.25	0.059
	p	*0.12*	*0.80*	*0.23*	*0.78*
Immediate	*r*	-0.207	-0.018	-0.142	0.359
	p	*0.30*	*0.93*	*0.50*	*0.08*
Delayed	*r*	-0.329	-0.034	-0.273	0.319
	p	*0.09*	*0.87*	*0.19*	*0.12*
Follow up	*r*	-0.303	-0.214	-0.157	-0.063
	p	*0.13*	*0.29*	*0.46*	*0.77*

Note: Significant correlations in bold. Density = spindle density (number of spindles per 30 seconds), Ampl. = average maximal spindle amplitude.

Continued on the next page

Correlations between cued recall of novel words and sleep spindle activity at each of the four electrodes.

		C3		C4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	-0.174	0.282	-0.203	0.329
	p	*0.42*	*0.18*	*0.31*	*0.09*
Immediate	*r*	**-0.443**	-0.025	-0.259	0.116
	p	***0.030***†	*0.91*	*0.19*	*0.56*
Delayed	*r*	**-0.466**	0.141	-0.338	0.285
	p	***0.022***†	*0.51*	*0.09*	*0.15*
Follow up	*r*	**-0.425**	-0.044	-0.238	-0.011
	p	***0.043***†	*0.84*	*0.25*	*0.96*

		F3		F4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	-0.121	0.236	0.122	0.123
	p	*0.55*	*0.24*	*0.56*	*0.56*
Immediate	*r*	-0.285	-0.104	-0.288	0.167
	p	*0.15*	*0.61*	*0.16*	*0.43*
Delayed	*r*	-0.304	0.029	-0.176	0.213
	p	*0.12*	*0.89*	*0.40*	*0.31*
Follow up	*r*	-0.257	-0.212	-0.174	-0.140
	p	*0.22*	*0.31*	*0.43*	*0.52*

Note: Significant correlations in bold. Density = spindle density (number of spindles per 30 seconds), Ampl. = average maximal spindle amplitude.

Continued on the next page

Correlations between novel word recognition RTs and sleep spindle activity at each of the four electrodes.

		C3		C4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	-0.040	-0.185	-0.185	-0.226
	p	*0.85*	*0.39*	*0.36*	*0.26*
Immediate	*r*	0.074	-0.058	-0.048	-0.044
	p	*0.73*	*0.79*	*0.81*	*0.83*
Delayed	*r*	0.158	-0.296	0.147	-0.347
	p	*0.46*	*0.16*	*0.47*	*0.08*
Follow up	*r*	0.278	-0.267	0.144	-0.205
	p	*0.20*	*0.22*	*0.48*	*0.32*

		F3		F4	
		Density	Ampl.	Density	Ampl.
Change overnight	*r*	0.070	-0.178	-0.044	0.091
	p	*0.73*	*0.37*	*0.82*	*0.67*
Immediate	*r*	0.071	0.014	-0.029	-0.313
	p	*0.72*	*0.95*	*0.89*	*0.14*
Delayed	*r*	0.194	-0.187	0.080	-0.325
	p	*0.33*	*0.35*	*0.70*	*0.11*
Follow up	*r*	0.298	-0.092	0.216	-0.385
	p	*0.14*	*0.65*	*0.31*	*0.06*

Note: Significant correlations in bold. Density = spindle density (number of spindles per 30 seconds), Ampl. = average maximal spindle amplitude

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439.

Axmacher, N., Draguhn, A., Elger, C. E., & Fell, J. (2009). Memory processes during sleep: Beyond the standard consolidation theory. *Cellular and Molecular Life Sciences, 66,* 2285-2297.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition : A review of the literature. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 187-222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bates, D. M. (2005). Fitting linear mixed models in R: Using the lme4 package. *R News: The Newsletter of the R Project, 5*, 27–30.

Bodner, G. E., & Masson, M. E. J. (2003). Beyond spreading activation: An influence of relatedness proportion on masked semantic priming. *Psychonomic Bulletin & Review, 10,* 645-652.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & Jada-Simone, S. W. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24,* 127-135.

Bourassa, D. C., & Besner, D. (1998). When do nonwords activate semantics? Implications for models of visual word recognition. *Memory & Cognition, 26,* 61-74.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Interfering neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition, 97,* B45-B54.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A-F., Sommer, J., Wolbers, T., & Knecht, S. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *NeuroImage, 25,* 958-968.

Breitenstein, C., & Knecht, S. (2002). Development of a language learning model for behavioural and functional-imaging studies. *Journal of Neuroscience Methods, 114,* 173-179.

Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime : Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience, 25,* 493-500.

Brown, G. D. A., & Lewandowsky, S. (in press). Forgetting in memory models: Arguments against trace decay and consolidation failure. In S. Della Salla (Ed.), *Forgetting*. Hove, UK: Psychology Press.

Bueno, S., & Frenck-Mestre, C. (2008). The activation of semantic memory: Effects of prime exposure, prime–target relationship, and task demands. *Memory & Cognition, 36,* 882-898.

Burns, D. J., & Gold, D. E. (1999). An analysis of item gains and losses in retroactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 978-985.

Buxton, C. E. (1943). The status of research in reminiscence. *Psychological Bulletin, 40,* 313-340.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, but not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences of the USA, 106,* 10130-10134.

Clay, F., Bowers, J. S., Davis, C. J., & Hanley, D. A. (2007). Teaching adults new words: the role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 970-976.

Clemens, Z., Fabo, D., & Halasz, P. (2005). Overnight verbal retention correlates with the number of sleep spindles. *Neuroscience, 132,* 529-535.

Clemens, Z., Fabo, D., & Halasz, P. (2006). Twenty-four hours retention of visuospatial memory correlates with the number of parietal sleep spindles. *Neuroscience Letters, 403,* 52-56.

Clemens, Z., Molle, M., Eross, L., Barsi, P., Halasz, P., & Born, J. (2007). Temporal coupling of parahippocampal ripples, sleep spindles and slow oscillations in humans. *Brain, 130,* 2868-2878.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32,* 193-210.

Connine, C. M., Titone, D., Deelman, T., & Blasko, D. (1997). Similarity mapping in spoken word recognition. *Journal of Memory and Language, 37,* 463-480.

Cornelissen, K., Laine, M., Renvall, K., Saarinen, T., Martin, N., & Salmelin, R. (2004). Learning new names for new objects: cortical effects as measured by magnetoencephalography. *Brain and Language, 89,* 617-622.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268-294.

Dagenbach, D., Carr, T. H., & Barnhardt, T. M. (1990). Inhibitory semantic priming of lexical decisions due to failure to retrieve weakly activated codes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 328-340.

Dagenbach, D., Horst, S., & Carr, T. H. (1990). Adding new information to semantic memory: How much learning is enough to produce automatic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 581-591.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128,* 165-185.

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21,* 803-820.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society, 364,* 3773-3800.

De Koninck, J., Lorrain, D., Christ, G., Proulx, G., & Coulombe, D. (1989). Intensive language learning and increases in rapid eye movement sleep: Evidence of a performance factor. *International Journal of Psychophysiology, 8,* 43-47.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods, 134,* 9-21.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11,* 114-126.

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews, 13,* 309-321.

Dobel, C., Junghofer, M., Breitenstein, C., Klauke, B., Knecht, S., Pantev, C., Zwitserlood, P. (in press). The adult brain acquires novel words in a fast, untutored and effortless way. *Journal of Cognitive Neuroscience.*

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General, 136,* 169-183.

Duff, M. C., Hengst, J., Tranel, D., & Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature Neuroscience, 9,* 140-146.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18,* 35-39.

Dumay, N., Gaskell, M. G., & Feng, X. (2004). A day in the life of a spoken word. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 339–344). Mahwah, NJ: Erlbaum.

Duyck, W., Desmet, T., Verbeke, L., & Brysbaert, M. (2004). WordGen: A tool for word selection and non-word generation in Dutch, German, English, and French. *Behavior Research Methods, Instruments & Computers, 36,* 488-499.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67,* 224-238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: stability over time. *Journal of the Acoustical Society of America, 119,* 1950-1953.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences of the USA, 104,* 7723-7728.

Ellenbogen, J. M., Hulbert, J. C., Jiang, Y., & Stickgold, R. (2009). The sleeping brain's influence on verbal memory: Boosting resistance to interference. *PLoS ONE, 4,* 1-4.

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: Sleep, declarative memory, and associative interference. *Current Biology, 16,* 1290-1294.

Ellenbogen, J. M., Payne, J. D., & Stickgold, R. (2006). The role of sleep in declarative memory consolidation: Passive, permissive, active, or none? *Current Opinion in Neurobiology, 16,* 716-722.

Ellis, A. W., Ferreira, R., Cathles-Hagan, P., Holt, K., Jarvis, L., & Barca, L. (2009). Word learning and the cerebral hemispheres: from serial to parallel processing of written words. *Philosophical Transactions of the Royal Society, 364,* 3675-3696.

Fenn, K. M., Gallo, D. A., Margoliash, D., Roediger, H. L., & Nusbaum, H. C. (2009). Reduced false memory after sleep. *Learning & Memory, 16,* 509-513.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425,* 614-616.

Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output : Lexicalization during artificial language learning. *Cognition, 112,* 349-366.

Ferrarelli, F., Huber, R., Peterson, M. J., Massimini, M., Murphy, M., Riedner, B. A., Watson, A., Bria, P., & Tononi, G. (2007). Reduced sleep spindle activity in schizophrenia patients. *American Journal of Psychiatry, 164,* 483-492.

Fogel, S. M., Smith, C. T. (2006). Learning-dependent changes in sleep spindles and stage 2 sleep. *Journal of Sleep Research, 15,* 250-255.

Fogel, S. M., Smith, C. T., & Cote, K. A., (2007). Dissociable learning-dependent changes in REM and non-REM sleep in declarative and procedural memory systems. *Behavioral Brain Research, 180,* 48-61.

Forster, K. I. (1985). Lexical acquisition and the modular lexicon. *Language and Cognitive Processes, 1,* 87-108.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 680-698.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116-124.

Gais, S., & Born, J. (2004a). Declarative memory consolidation: Mechanisms acting during human sleep. *Learning & Memory, 11,* 679-685.

Gais, S., & Born, J. (2004b). Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences of the USA, 101,* 2140-2144.

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory, 13,* 259-262.

Gais, S., Molle, M., Helms, K., & Born, J. (2002). Learning-dependent increases in sleep spindle density. *The Journal of Neuroscience, 22,* 6830-6834.

Gais, S., Plihal, W., Wagner, U., & Born, J. (2000). Early sleep triggers memory for early visual discrimination skills. *Nature Neuroscience, 3,* 1335-1339.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89,* 105-132.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166-1183.

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1122-1139.

Gomez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science, 17,* 670-674.

Gronholm, P., Rinne, J. O., Vorobyev, V., Laine, M. (2005). Naming of newly learned object: A PET activation study. *Cognitive Brain Research, 25,* 359-371.

Gronholm, P., Rinne, J. O., Vorobyev, V., Laine, M. (2007). Neural correlates of naming newly learned objects in MCI. *Neuropsychologia, 45,* 2355-2369.

Grossi, G. (2006). Relatedness proportion effects on masked associative priming: An ERP study. *Psychophysiology, 43,* 21-30.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology, 56A,* 1213-1236.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 106,* 491-528.

Hasher, L., Goldstein, D., & May, C. P. (2005). It's about time: Circadian rhythms, memory, and aging. In C. Izawa, & N. Ohta (Eds.), *Human Learning and Memory: Advances in Theory and Application*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W.C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology 10,* 431–436.

Horne, J. (2006). *Sleepfaring: A journey through the science of sleep.* Oxford: Oxford University Press.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression.* New York: Wiley Inter-Science.

Hupbach, A., Gomez, R. L., Bootzin, R. R., Nadel, L. (2009). Nap-dependent learning in infants. *Developmental Science, 12,* 1007-1012.

Hutchison, K. A. (2003). Is semantic priming due to association strength or featural overlap? A micro-analytic review. *Psychonomic Bulletin & Review, 10,* 785-813.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446.

James, T. W., & Gauthier, I. (2004). Brain areas engaged during visual judgements by involuntary access to novel semantic information. *Vision Research, 44,* 429-439.

Jarrold, C., & Thorn, A. (2007, July). *Developmental changes in lexical and sublexical influences on new word learning.* Poster presented at the EPS/Psychonomic Society meeting, Edinburgh, UK.

Johnston, M., McKague, M., & Pratt, C. (2004). Evidence for an automatic orthographic code in the processing of visually novel word forms. *Language and Cognitive Processes, 19,* 273-317.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society. B: Biological Sciences, 362,* 857-875.

Kouider, S., & Dupoux, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. *Psychological Science, 15,* 75-81.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51,* 141-178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13,* 262-268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56,* 1-15.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19,* 332-338.

Kuriyama, K., Stickgold, R., Walker, M. P. (2004). Sleep-dependent learning and motor skill complexity. *Learning & Memory, 11,* 705-713.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203-205.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research, 17,* 3-10.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55,* 306-353.

Lechner, H. A., Squire, L. R., & Byrne, J. H. (1999). 100 years of consolidation – Remembering Muller and Pilzecker. *Learning & Memory, 6,* 77-87.

Litman, L., & Davachi, L. (2008). Distributed learning enhances relational memory consolidation. *Learning & Memory, 15,* 711-716.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7,* 618-630.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. *Journal of Experimental Psychology: General, 132,* 202-227.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London: Series B, Biological Sciences, 176,* 161-234.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 262,* 23-81.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences, 11,* 442-450.

Marshall, L., Helgadottir, H., Molle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature, 444,* 610-613.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

Mattys, S. L., & Clark, J. H. (2002). Lexical activity in speech processing: evidence from pause detection. *Journal of Memory and Language, 47,* 343-359.

McCandliss, B. D., Posner, M. I., & Givon, T. (1997). Brain plasticity in learning visual words. *Cognitive Psychology, 33,* 88-110.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 10,* 1-86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995), Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419-457.

McKague, M., Pratt, C., & Johnston, M. B. (2001). The effect of oral vocabulary on reading visually novel words: a comparison of the dual-route-cascaded and triangle frameworks. *Cognition, 80,* 231-262.

McKay, A., Davis, C., Savage, G., & Castles, A. (2008). Semantic involvement in reading aloud: Evidence from a nonword training study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1495-1517.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience, 7*, 703-704.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science, 317*, 631-631.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* New York: Psychology Press.

McQueen, J. M. (2007). Eight Questions About Spoken Word Recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 37-53). Oxford: Oxford University Press.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113-1126.

Meeter, M., & Murre, J. M. J. (2004). Consolidation of long-term memory: Evidence and alternatives. *Psychological Bulletin, 130,* 843-857.

Meeter, M., & Murre, J. M. J. (2005). TraceLink: A model of consolidation and amnesia. *Cognitive Neuropsychology, 22,* 559-587.

Meier-Koll, A., Bussmann, B., Schmidt, C., & Neuschwander, D. (1999). Walking through a maze alter the architecture of sleep. *Perceptual and Motor Skills, 88,* 1141-1159.

Mestres-Misse, A., Camara, E., Rodriguez-Fornells, A., Rotte, M., & Munte, T. F. (2008). Functional neuroanatomy of meaning acquisition from context. *Journal of Cognitive Neuroscience, 20,* 2153-2166.

Mestres-Misse, A., Rodriguez-Fornells, A., & Munte, T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex, 17,* 1858-1866.

Milner, C. E., Fogel, S. M., & Cote, K. A. (2006). Habitual napping moderates motor performance improvements following a short daytime nap. *Biological Psychology, 73,* 141-156.

Monsell, S. (1985). Repetition and the lexicon. In A. W. Ellis (Ed.), *Progress in the psychology of language, Volume 2* (pp. 147-195). Hillsdale, NJ: Lawrence Erlbaum Associates.

Morin, A., Doyon, J., Dostie, V., Barakat, M., Tahar, A. H., Korman, M., Benali, H., Karni, A., Ungerleider, L., Carrier, J. (2008). Motor sequence learning increses sleep spindles and fast frequencies in post-training sleep. *Sleep, 31,* 1149-1156.

Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A., & Rosenbaum, R. S. (2006). The cognitive neuroscience of remote episodic, semantic, and spatial memory. *Current Opinion in Neurobiology, 16,* 179-190.

Mueller, G. E., & Pilzecker, A. (1900). Experimentelle Beitrage zur Lehre vom Gedachtniss [Experimental contributions to the science of memory]. *Zeitschrift fuer Psychologie, 1,* 1–288.

Musen, G., & Squire, L. R. (1991). Normal acquisition of novel verbal information in amnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1095-1104.

Nader, R., & Smith, C. (2003). A role for stage 2 sleep in memory processing. In P. Maquet, C. Smith & R. Stickgold (Eds.), *Sleep and brain plasticity* (pp. 87-98). Oxford: Oxford University Press.

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: effects of exposure, durability, and context. *Journal of Experimental Child Psychology, 96,* 71-84.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: two modes of word acquisition? *Developmental Science, 6,* 136-142.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Neely, J. H., & Keefe, D. E. (1989). Semantic context effects in visual word processing: A hybrid prospective-retrospective processing theory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24, pp. 207-248). New York: Academic Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407.

Nelson, J. R., Balass, M., & Perfetti, C. A. (2005). Differences between written and spoken input in learning new words. *Written Language & Literacy, 8,* 25-44.

Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PLoS ONE, 2,* 341-347.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition, 52,* 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204-238.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443-512.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L-W., Wamsley, E. J., Tucker, M. A., Walker, M. P., & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory, 92,* 327-334.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Phillips, C., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A., & Maquet, P. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron, 44,* 535-545.

Peigneux, P., Orban, P., Balteau, E., Degueldre, C., Luxen, A., Laureys, S., & Maquet, P. (2006). Offline persistence of memory-related cerebral activity during active wakefulness. *PLoS Biology, 4,* 647-658.

Perea, M., Dunabeitia, J. A., & Carreiras, M. (2008). Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory and Language, 58,* 916-930.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition, 62,* 223-240.

Perea, M., & Rosa, E. (2002). Does the proportion of associatively related pairs modulate the associative priming effect at very brief stimulus-onset asynchronies? *Acta Psychologica, 110,* 103-124.

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1281-1292.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107,* 786-823.

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9,* 534-547.

Plihal, W., & Born, J. (1999). Effects of early and late nocturnal sleep on priming and spatial memory. *Psychophysiology, 36,* 571-582.

Polster, M. R., Nadel, L., & Schacter, D. L. (1991). Cognitive neuroscience analyses of memory: A historical perspective. *Journal of Cognitive Neuroscience, 3,* 95-116.

Potts, G. R., St. John, M. F., & Kirson, D. (1989). Incorporating new information into existing world knowledge. *Cognitive Psychology, 21,* 303-333.

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods, 39,* 859-862.

Qiao, X, Forster, K., & Witzel, N. (2009). Is banara really a word? *Cognition, 113,* 254-257.

R Development Core Team. (2007). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org

Rajaram, S., & Neely, J. H. (1992). Dissociative masked repetition priming and word frequency effects in lexical decision and episodic recognition tasks. *Journal of Memory and Language, 31,* 152-182.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes, 15,* 507-537.

Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized terminology, techniques, and scoring system for sleep stage scoring of human subjects.* Bethesda, MD: U. S. Department of Health, Education and Welfare.

Ribot, T. (1882). *Diseases of memory.* New York: Appleton.

Rogers, T. T., & Mayberry, E. Sleep influences acquisition of new semantic information. *Manuscript in preparation.*

Roodenrys, S., & Hinton, M. (2002). Sublexical or lexical effects on serial recall of nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 29-33.

Rueckl, J. G., & Olds, E. M. (1993). When pseudowords acquire meaning: Effect of semantic associations on pseudoword repetition priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 515-527.

Rueckl, J. G., & Dror, I. E. (1994). The effect of orthographic-semantic systematicity on the acquisition of new words. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV*. London: Erlbaum.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language, 35,* 606-621.

Salasoo, A., Shiffrin, R. M., & Feustel, T. C. (1985). Building permanent memory codes: codification and repetition effects in word identification. *Journal of Experimental Psychology: General, 114,* 50-77.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71,* 1207-1218.

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience, 5*, 700-703.

Schabus, M., Dang-Vu, T. T., Albouy, G., Balteau, E., Boly, M., Carrier, J., Darsaud, A., Degueldre, C., Desseilles, M., Gais, S., Phillips, C., Rauchs, G., Schnakers, C., Sterpenich, V., Vandewalle, G., Luxen, A., & Maquet, P. (2007). Hemodynamic cerebral correlates of sleep spindles during human non-rapid eye movement sleep. *Proceedings of the National Academy of Sciences of the USA, 104,* 13164-13169.

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Kloesch, G., Anderer, P., Klimesch, W., Saletu, B., Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep, 27,* 1479-1485.

Schabus, M., Hoedlmoser, K., Pecherstorfer, T., Anderer, P., Gruber, G., Parapatics, S., Sauter, C., Kloesch, G., Klimesch, W., Saletu, B., & Zeitlhofer, J. (2008). Interindividual sleep spindle differences and their relation to learning-related enhancements. *Brain Research, 1191,* 127-135.

Schmidt, C., Peigneux, P., Muto, V., Schenkel, M., Knoblauch, V., Munch, M., de Quervain, D. J. F., Wirtz-Justice, A., & Cajochen, C. (2006). Encoding

difficulty promotes postlearning changes in sleep spindle activity during napping. *The Journal of Neuroscience, 26,* 8976-8982.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20,* 11-21.

Sereno, J. A. (1991). Graphemic, associative, and syntactic priming effects at a brief stimulus onset asynchrony in lexical decision and naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 459-477.

Siapas, A. G., & Wilson, M. A. (1998). Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron, 21,* 1123-1128.

Siegel, J. M. (2001). The REM sleep - memory consolidation hypothesis. *Science, 294,* 1058-1063.

Sirota, A., Csicsvari, J., Buhl, D., & Buzsaki, G. (2003). Communication between neocortex and hippocampus during sleep in rodents. *Proceedings of the National Academy of Sciences of the USA, 100,* 2065-2069.

Slowiaczek, L. M. (1994). Priming in a single-word shadowing task. *The American Journal of Psychology, 107,* 245-260.

Snoeren, N. D., Gaskell, M. G., & Di Betta, A. M. (2009). The perception of assimilation in newly learned novel words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 542-549.

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10,* 281-284.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195-231.

St. Clair, M. C., & Monaghan, P. (2008). Language abstraction: Consolidation of language structure during sleep. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Stickgold, R. (2009). How do I remember? Let me count the ways. *Sleep Medicine Reviews, 13,* 305-308.

Stickgold, R., James, L., & Hobson, J. A. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience, 3,* 1237-1238.

Stickgold, R., Scott, L., Rittenhouse, C., & Hobson, J. A. (1999). Sleep-induced changes in associative memory. *Journal of Cognitive Neuroscience, 11,* 182-193.

Stolz, J. A., & Besner, D. (1996). Role of set in visual word recognition: Activation and activation blocking as nonautomatic processes. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1166-1177.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44,* 1321-1337.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25,* 201-221.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language, 36,* 291-321.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49,* 1175-1192.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54,* 99-132.

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L., Rijpkema, M., & Fernandez, G. (2009). Shift from hippocampal to neocortical retrieval network with consolidation. *The Journal of Neuroscience, 29,* 10087-10093.

Takehara-Nishiuchi, K., & McNaughton, B. L. (2008). Spontaneous changes of neocortical code for associative memory during consolidation. *Science, 322,* 960-963.

Tamaki, M., Matsuoka, T., Nittono, H., & Hori, T. (2008). Fast sleep spindle (13-15 Hz) activity correlates with sleep-dependent improvement in visuomotor performance. *Sleep, 31,* 204-211.

Tamaki, M., Matsuoka, T., Nittono, H., & Hori, T. (2009). Activation of fast sleep spindles at the premotor cortex and parietal areas contributes to motor learning: A study using sLORETA. *Clinical Neurophysiology, 120,* 878-886.

Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *Quarterly Journal of Experimental Psychology, 61,* 361-371.

Thorn, A. S. C., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 729-735.

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Reviews, 10,* 49-62.

Tucker, M. A., & Fishbein W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep, 31,* 197-203.

Underwood, B. J. (1945). The effect of successive interpolations on retroactive and proactive inhibition. *Psychological Monographs, 59.*

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38,* 584-589.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135,* 452-477.

Van Ormer, E.B. (1933). Sleep and retention. *Psychological Bulletin, 30,* 415 - 439.

Vertes, R. P. (2004). Memory consolidation in sleep: Dream or reality? *Neuron, 44,* 135-148.

Vertes, R. P. & Eastman, K. E. (2000). The case against memory consolidation in REM sleep. *Behavioral and Brain Sciences, 23,* 867-876.

Vertes, R. P., & Siegel, J. M., (2005). Time for the sleep community to take a critical look at the purported role of sleep in memory processing. *Sleep, 28,* 1228-1229.

Wagner, U., Gais, S., & Born, J. (2001). Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learning & Memory, 8,* 112-119.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature, 427,* 352-355.

Walker, M. P. (2005). A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences, 28,* 51-104.

Walker, M. P. (2009). The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences, 1156,* 168-197.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron, 35,* 205-211.

Walker, M. P., Brakefield, T., Seidman, J., Morgan, A., Hobson, J. A., & Stickgold, R. (2003). Sleep and time course of motor skill learning. *Learning & Memory, 10,* 275-284.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron, 44,* 121-133.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psychology, 57,* 139-166.

Whittlesea, B. W. A., & Cantwell, A. L. (1987). Enduring influence of the purpose of experiences: Encoding-retrieval interactions in word and pseudoword perception. *Memory & Cognition, 15,* 465-472.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265,* 676-679.

Wirth, S., Yanike, M., Frank, L. M., Smith, A. C., Brown, E. N., & Suzuki, W. A. (2003). Single neurons in the monkey hippocampus and learning of new associations. *Science, 300,* 1578-1581.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55,* 235-269.

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science, 14,* 6-9.

Zola-Morgan, S. M., & Squire, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science, 250,* 288-290.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32,* 25–64.